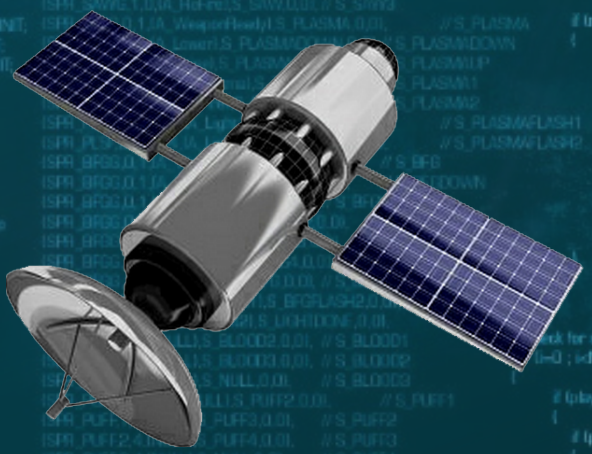
ASTES

Advances in Science, Technology & Engineering Systems Journal



VOLUME 3-ISSUE 1|JAN-FEB 2018

www.astesj.com
ISSN: 2415-6698

Editorial

Advances in Science, Technology and Engineering Systems Journal (ASTESJ) is an online-only journal dedicated to publishing significant advances covering all aspects of technology relevant to the physical science and engineering communities. The journal regularly publishes articles covering specific topics of interest.

Current Issue features key papers related to multidisciplinary domains involving complex system stemming from numerous disciplines; this is exactly how this journal differs from other interdisciplinary and multidisciplinary engineering journals. This issue contains 19 accepted papers in Electrical and Information Technology domains.

Editor-in-chief
Prof. Passerini Kazmersk

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Volume 3 Issue 1 **January-February 2018**

CONTENTS

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 01-10 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Machine Learning framework for image classification

Sehla Loussaief*,[1,2], Afef Abdelkrim[1,2]

[1] LA.R.A, Ecole Nationale d'Ingénieurs de Tunis, Université Tunis El Manar. BP 32, le Belvédère 1002, Tunisia

[2] ENICarthage, Université de Carthage, 35 rue des Entrepreneurs, Charguia II, Tunis, Tunisia

ARTICLE INFO

Article history:
Received: 27 October, 2017
Accepted: 10 December, 2017
Online: 18 January, 2018

Keywords :
Image classification
Features extraction
Bag of Features
Class prediction accuracy
Speed Up Robust Features
Machine learning

ABSTRACT

Hereby in this paper, we are going to refer image classification. The main issue in image classification is features extraction and image vector representation. We expose the Bag of Features method used to find image representation. Class prediction accuracy of varying classifiers algorithms is measured on Caltech 101 images. For feature extraction functions we evaluate the use of the classical Speed Up Robust Features technique against global color feature extraction. The purpose of our work is to guess the best machine learning framework techniques to recognize the stop sign images. The trained model will be integrated into a robotic system in a future work.

1. Introduction

This paper is an extension of work originally presented in the 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2016. It presents the use of machine learning algorithms for image classification and exposes the Bag of Features (BoF) approach. The BoF aims at finding vector representations of input images that can be used to categorize images into a finite set of classes. This paper attempts to give a comparison between different features extraction methods and classification algorithms.

The remainder of this paper is structured as follows: Section 2 provides background information on machine learning. Section 3 gives a brief review of computer vision system, while Section 4 provides a detailed description of the Bag of Features paradigm. It also exposes the Speed Up Robust Features (SURF) detector of image Region Of Interest (ROI) and highlights the unsupervised K-Means algorithm. In section 5 we describe different learning algorithms that we will use as classifiers. Section 6 discusses experimentations carried out in order to evaluate the classification accuracy of our machine learning framework in Caltech 101 image dataset. We conclude with a discussion of open questions and current direction of image classification and feature extraction research.

2. Machine Learning Paradigm

In the past years, computer scientists developed a wide variety of algorithms particularly suited to prediction. Among these we cite: Nearest Neighbor Classification, Neural Nets, Ensembles of Trees and Support Vector Machines. These machine learning (ML) methods are easier to implement and perform better than the classical statistical approaches.

Statistical approaches to model fitting, which have been the standard for decades, start by assuming an appropriate data model witch parameters are then estimated from the data. By contrast, ML avoids starting with a data model and rather uses an algorithm to learn the relationship between the response and its predictors. The statistical approach focuses on issues such as what model will be postulated how the response is distributed, and whether observations are independent. By contrast, the ML approach assumes that the data-generating process is complex and unknown, and tries to learn the response by observing inputs and responses and finding dominant patterns [1-2].

The machine learning workflow is described in Figure 1:

*Sehla Loussaief, sehla.loussaief@gmail.com

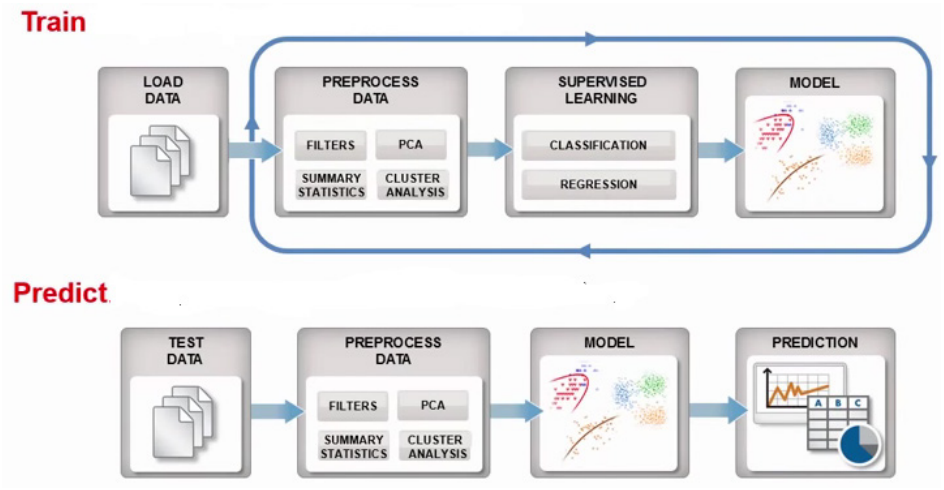


Figure 1: Machine learning workflow

Machine Learning focuses on what is being predicted, the model's ability to predict well and how to measure prediction success.

Many fields of modern society use Machine-learning technologies: web searches, content filtering on social net-works, recommendations on e-commerce websites. Today, ML is present in consumer products such as cameras and smartphones. Machine-learning systems are used in computer vision, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results of search.

3. Computer Vision System

Computer Vision System provides algorithms, functions, and applications for designing and simulating computer vision and video processing systems. It offers image classification and retrieval [3–6], object recognition and matching [7-9], 3D scene reconstruction [10], robot localization [11], object detection and tracking and video processing. All of these processing systems rely on the presence of stable and meaningful features in the image. Thus, the most important steps in these applications are detecting and extracting the image features.

The approach consists in detecting interest regions (key-points) in each image that are covariant to a class of transformations. Then, for each detected regions, an invariant feature vector representation (i.e., descriptor) for image data around the detected key-points is built.

Two types of image features can be extracted for image content representation; namely global features and local features. Global features (e.g., color and texture) describe an image as a whole. While, local features aim to detect key-points or interest regions in an image and describe them. In this context, if the local feature algorithm detects *n* key-points in the image, there are *n* vectors describing each one's shape, color, orientation, texture and more.

The use of global color and texture features is an efficient technic for finding similar images in a dataset. While the local structure oriented features are adequate for object classification. It was proven that using global features cannot distinguish foreground from background of an image, and mix information from both parts together. [12].

In this work we deploy and test a machine learning based framework in object detection and recognition. We are interested on image category classification. To achieve tests we use the Calltech[1] dataset.

As the main issue in image classification is image features extraction, we use in our research the Bag of Features (BoF) techniques described in section 4.

4. Bag of Features Paradigm for Image Classification

In document classification fields (text documents), a bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, the bag-of-words model (BoW model) can be applied to image classification, by treating image features as words.

In computer vision, a bag of visual words is a vector of occurrence counts of a vocabulary of local image features. To encode an image using BoW model, an image can be treated as a document. Thus, "words" in images need to be defined. For this purpose, we use three steps: feature detection, feature description, and codebook generation [13-15].

4.1. Features Detection

In image processing the concept of feature detection refers to techniques that aim at abstractions of image information. Computer vision is using these extracted informations in making local decisions. Given that, a feature is defined as an "interesting" part of an image. The resulting features will be subsets of the image domain, often in the form of isolated points, continuous curves or connected regions [16].

Feature detection is a low-level image processing operation. That is, it is usually performed as the first operation on an image. It examines every pixel to see if there is a feature present at that

[1] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

pixel. If this is part of a larger algorithm, then the algorithm will typically only examine the image in the region of the features.

As a built-in pre-requisite to feature detection, the input image is usually smoothed by a Gaussian kernel in a scale-space representation and one or several feature images are computed, often expressed in terms of local image derivatives operations [17].

Common features detectors: Canny, Sobel, Level curve curvature, FAST, Laplacian of Gaussian, MSER, Grey-level blobs.

4.2. Features Description

After feature detection, each image is abstracted by several local patches. Feature representation methods represent the patches as numerical vectors called feature descriptors. A descriptor should have the ability to handle intensity, rotation, scale and affine variations to some extent.

One of the most famous descriptors is Scale-invariant feature transform (SIFT) [18]. SIFT converts each patch to 128-dimensional vector. After this step, each image is a collection of vectors of the same dimension (128 for SIFT), where the order of different vectors is of no importance.

4.3. Codebook Generation

Finally, the BoW model converts vector-represented patches to "codewords", which also produces a "codebook" (word dictionary). A codewords can be considered as a representative of several similar patches.

The most used method for building a codebook is performing k-means (section 4.5) clustering over all the vectors. Codewords are then defined as the centers of the learned clusters. The number of the clusters is the codebook size (the size of the word dictionary). Thus, each patch in an image is mapped to a certain codeword through the clustering process and the image can be represented by the histogram of the codewords [19].

In image classification, an image is classified according to its visual content. The feature vector consists of SIFT/SURF features computed on a regular grid across the image and vector quantized into visual words.

The frequency of each visual word is then recorded in a histogram for each tile of a spatial tiling.

The final feature vector for the image is a concatenation of these histograms.

4.4. Speed Up Robust Features (SURF) Extraction Technique

The Speed Up Robust Features method extracts salient features and descriptors from images. This extractor is preferred over Scale-Invariant Feature Transform (SIFT) due to its concise descriptor length. The standard SIFT implementation uses a descriptor consisting of 128 floating point values.

SURF algorithm condenses this descriptor length to 64 floating point values.

It constructs a descriptor vector of length 64 using a histogram of gradient orientations in the local neighborhood around each key-point [20].

SURF considers the processing of grey-level images only, since they contain enough information to perform feature extraction and image analysis [21].

The implementation of SURF used in this paper is provided by the Matlab R2015a library.

4.5. Descriptors clustering: K-Means

Bag of Features (BoF) model is a key development in image classification using key-points and descriptors.

The descriptors extracted from the training images are grouped into N clusters of visual words using unsupervised learning algorithms such as K-means. A descriptor is categorized into its cluster centroid using an "Euclidean distance" metric. For input image, each extracted descriptor is mapped into its nearest cluster centroid.

A histogram of counts is constructed by incrementing a cluster centroid's number of occupants each time a descriptor is placed into it. The result is that each image is represented by a histogram vector of length N. It is necessary to normalize each histogram by its L2-norm to make this procedure invariant to the number of descriptors used. Applying Laplacian smoothing to the histogram appears to improve classification results.

K-means clustering is selected over Expectation Maximization (EM) to group the descriptors into N visual words [22]. Experimental methods verify the computational efficiency of K-means as opposed to EM. Our specific application necessitates rapid training which precludes the use of the slower EM algorithm.

5. Learning and Recognition Based on BoW Models

Computer vision researchers have developed several learning methods to leverage the BoF model for image related tasks. For multiple label categorization problems, the confusion matrix can be used as an evaluation metric.

A confusion matrix is defined as a specific table layout that allows visualization of the performance of a supervised learning algorithm. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice-versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another) [23].

In this work we investigate many supervised learning algorithms. These learning algorithms are used to classify an image using the histogram vector previously constructed in the K-means step.

5.1. Support Vector Machine

Classifying data is a common task in machine learning. A support vector machine (SVM) technic constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space.

It can be used for classification, regression, or other tasks. A good separation is achieved by the hyperplane that has the longest distance to the nearest training-data point of any class (so-called functional margin). A larger margin induces lower generalization error of the classifier [24].

Classification of images and Hand-written character recognition can be performed using SVMs. The SVM algorithm has, also, been widely applied in the biological and other sciences.

5.2. Nearest Neighbor Classification

In pattern classification, the k-nearest neighbors (kNN) rule is the oldest. It is also considered as the simplest methods. The kNN rule classifies each unlabeled example by the majority label among its k-nearest neighbors in the training set. The distance metric used to identify nearest neighbors influences greatly its overall performance.

In the absence of prior knowledge, most kNN classifiers use simple Euclidean distances to measure the dissimilarities between examples represented as vector inputs.

Euclidean distance metrics, however, do not capitalize on any statistical regularity in the data that might be estimated from a large training set of labeled examples [25].

5.3. Boosted Regression Trees (BRT)

The BRT technique target is to improve the performance of a single model. This is achieved by fitting many models and combining them for prediction. BRT uses two algorithm categories:

- Decision tree learning algorithm which uses a decision tree as a predictive model. It maps observations about an item to conclusions about its target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. When the target variable can take a finite set of values, the tree models are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees [26].
- Gradient boosting algorithm which is a machine learning technique used for regression and classification problems. It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion and it generalizes them by allowing optimization of an arbitrary differentiable loss function [27].

6. Experiments and Evaluation

This section provides an overview of different experiments that we use to evaluate the performance of our image classification machine learning framework.

We next describe the dataset used for testing followed by experimentation of SURF local features extractor. Next, we evaluate the impact of the categories number used in training on the accuracy of prediction. The last part of our work will focus on comparing the accuracy of different classifiers.

6.1. Dataset

Our results are reported on Calltech 101 image dataset to which we have added some new images of existing categories. Pictures of objects belong to 101 categories. Each category includes 40 to 800 images. The dataset was collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato. The size of each image is roughly 300 x 200 pixels. We are interested in stop sign category recognition.

6.2. SURF Local Feature Extractor and Descriptor

In the first part of experimentation we test the local feature extractor SURF and its robustness in matching features even after rotation and scaling image.

SURF is a scale and rotation invariant interest point detector and descriptor.

The feature finding process is usually composed of two steps; first, find the interest points in the image which might contain meaningful structures; this is usually done by comparing the Difference of Gaussian (DoG) in each location in the image under different scales. The second step is to construct the scale invariant descriptor on each interest point found in the previous step.

As first experimentation we use Matlab functionalities to test the SURF point of interest extraction function on sign stop images (Figure. 2, Figure. 3). Next we test the SURF matching features capability (Figure. 4).

6.3. Bag of Features Image Encoding

Features extracted in the first step will be used to represent each image category. To do that, the K-means clustering is used to reduce the number of features for proper classification. Only strongest features are considered. The encode approach is then applied. Thus, each image of the dataset is encoded into a vector feature using BoF.

The feature vector of an image represents the histogram of visual word occurrences contained in it. This histogram considered a basis for training the classifier. Figure 5 represents encoding results for some stop sign images.

6.4. Classifier Training Process

The encoded training images from each category are fed into a classifier training process to generate a predictive model.

Figure 6 illustrates the steps of the approach used in our image classification framework.

Figure 2: SURF Features Detection

Figure 3: SURF Features Detection in rotated (30°) and scaled (1.5) image

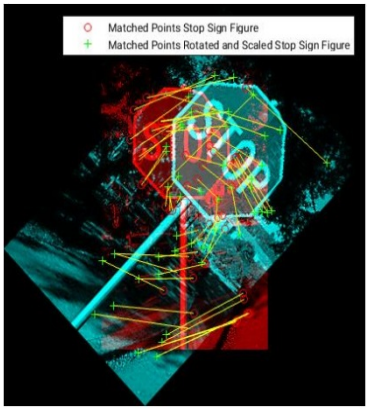
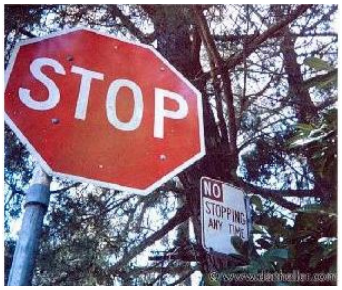
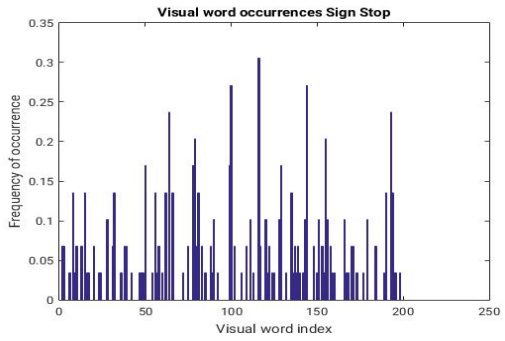


Figure 4: SURF point matching capabilities





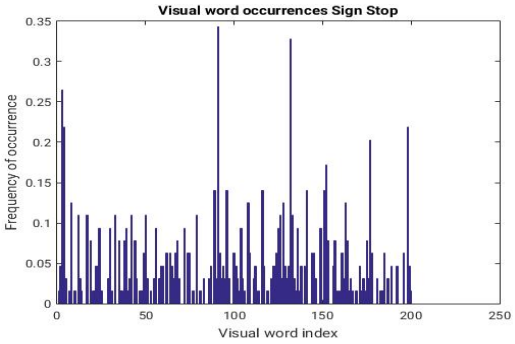


Figure 5: Histogram of visual words occurrences on stop sign images

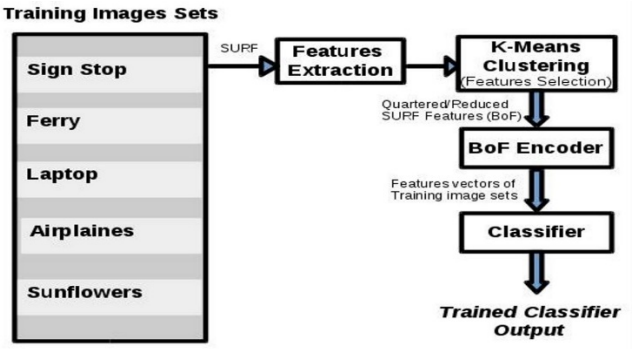


Figure 6: Image classification process

through different experimentations. For this purpose, we use some image categories from the Calltech101dataset. These classes are described in Table 1.

Experiment 1: Two categories classification accuracy measurement using SURF extractor

The Linear SVM classifier is used to generate a prediction model based on two image classes. During the training process we use 70% of the whole image dataset. The remaining images are included in the test dataset and used in the prediction assessment step. Measurements report that the achieved prediction average accuracy is 0.99 (Table 2).

Experiment 2: Three categories classification accuracy measurement using SURF extractor

During this experiment we fix the classifier to Linear SVM and the number of image categories to three. Image dataset is split to training dataset (70% of image dataset) and test dataset.

Measurements show that the achieved prediction average accuracy is 0.89 (Table 3).

Experiment 3: Four categories classification accuracy measurement using SURF extractor

For this prediction accuracy evaluation, we use the Linear SVM classifier and increase the image categories to four in order to measure the average accuracy of the classification process. It is reported that this achieved average accuracy is 0.88 (Table4).

Experiment 4: Five categories classification accuracy measurement using SURF extractor.

The Linear SVM classifier is used to classify between five image categories. It is reported that the achieved prediction average accuracy is 0.78 (Table 5).

As shown in Figure 7, we notice that the average accuracy of the classifier is influenced by the number of categories in training dataset. This metric is lower when the numbers of sets increase.

Experiment 5: Three categories classification accuracy measurement using a custom features extraction function: Color extractor.

During this experiment we use the Linear SVM as classifier and fix the number of image categories to three. For image vector representation we use a global features extractor instead of the SURF technique.

It is reported that the prediction average accuracy is 0.76 (Table 6) which is lower than the one achieved during Experiment 2.

Table 1: Image Dataset categories

Set category	Stop Sign Images	Ferry Images	Laptop images	Airplanes images	Sunflower images
Set size	69	67	81	800	85

Table 2: Learning confusion matrix with two image categories

	Predicted	
Known	Stop Sign	Ferry
Stop Sign	0.98	0.02
Ferry	0.00	1.00

Table 3: Learning confusion matrix with three image categories

	Predicted		
Known	Stop Sign	Laptop	Ferry
Stop Sign	0.93	0.04	0.03
Laptop	0.02	0.77	0.21
Ferry	0.00	0.02	0.98

Table 4: Learning confusion matrix with four image categories

	Predicted			
Known	Stop Sign	Laptop	Ferry	Airplanes
Stop Sign	0.95	0.05	0.00	0.00
Laptop	0.00	0.90	0.10	0.00
Ferry	0.00	0.00	0.85	0.15
Airplanes	0.00	0.05	0.10	0.85

Table 5: Learning confusion matrix with five image categories

	Predicted				
Known	Stop Sign	Laptop	Ferry	Airplanes	Sunflowers
Stop Sign	0.97	0.00	0.00	0.00	0.03
Laptop	0.07	0.49	0.14	0.10	0.20
Ferry	0.02	0.00	0.76	0.17	0.05
Airplanes	0.00	0.04	0.20	0.75	0.01
Sunflowers	0.00	0.00	0.02	0.00	0.98

Table 6: Learning confusion matrix using global color features extractor

	Predicted		
Known	Stop Sign	Laptop	Ferry
Stop Sign	0.84	0.04	0.12
Laptop	0.10	0.67	0.23
Ferry	0.02	0.22	0.76

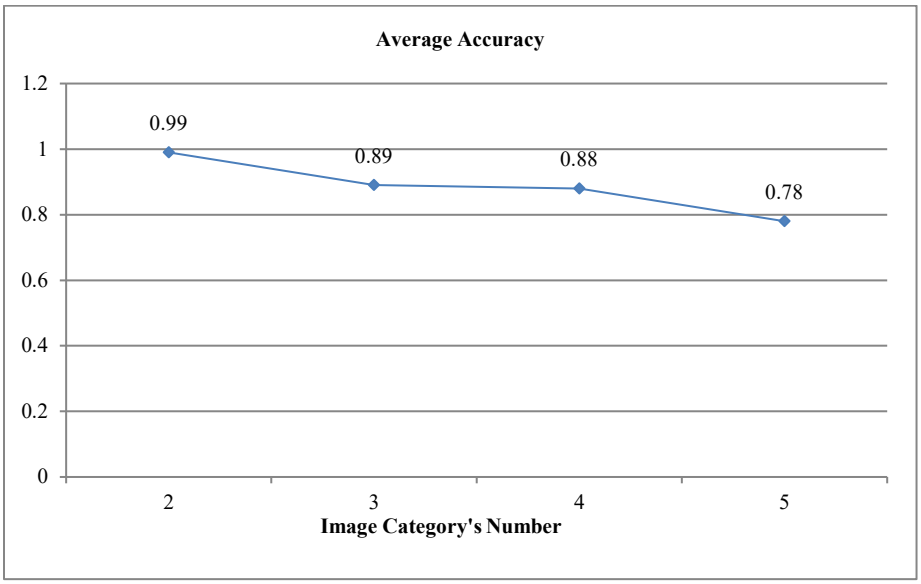


Figure: 7: Average accuracy variation based on image category's number

We notice that in our approach is better to use a Local feature extractor (SURF) than a global features extractor. This result is expected as the global features extraction technique is better with scene categorization and examination of surrounding environment (an image may be categorized as an office, forest, sea or street image) and not for object classification [28].

Experiment 6: Evaluating image category classification using different training learner.

We next fix the number of categories to 4, the features extraction technique to SURF and evaluate prediction models on varying the classifier algorithm. In each test we calculate the confusion matrix and average accuracy on validation dataset. We use 70% of data as a training set. The application trains the model on training set and assesses the performance with the validation set. Tables 7 to Table 14 illustrate the obtained confusion matrix.

We then generate the histogram (Figure 8) of the average accuracy based on the training classifier. For this purpose we varied the classifier trainer from SVM, KNN and ensemble classifier categories.

Measurements show that the image classification process performs better when we use a likehood SVM. It's reported that the Cubic SVM yields average accuracy which reaches 90%. The KNN techniques offer an average accuracy around 65%. Among the ensemble classifier trainers (2 last tested algorithms) the bagged trees achieves the best accuracy.

Table 7: Confusion matrix for Linear SVM

	Airplane	Ferry	Laptop	Stop sign
Airplane	17 85.5%	2 10.0%	1 5.0%	0 0.0%
Ferry	3 15.0%	17 85.0%	0 0.0%	0 0.0%
Laptop	0 0.0%	2 10.0%	18 90.0%	0 0.0%
Stop sign	0 0.0%	0 0.0%	1 5.0%	19 95.0%

Table 8: Confusion matrix for Fine Gaussian SVM

	Airplane	Ferry	Laptop	Stop sign
Airplane	0 0.0%	0 0.0%	20 10.0%	0 0.0%
Ferry	0 0.0%	0 0.0%	20 100.0%	0 0.0%
Laptop	0 0.0%	0 0.0%	20 100.0%	0 0.0%
Stop Sign	0 0.0%	0 0.0%	18 90.0%	2 10.0%

Table 9: Confusion matrix for Quadratic SVM

	Airplane	Ferry	Laptop	Stop Sign
Airplane	16 80.0%	3 15.0%	1 5.0%	0 0.0%
Ferry	3 15.0%	17 85.0%	0 0.0%	0 0.0%
Laptop	0 0.0%	1 5.0%	19 95.0%	0 0.0%
Stop Sign	0 0.0%	0 0.0%	5 25.0%	15 75.0%

Table 10: Confusion matrix for Cubic SVM

	Airplane	Ferry	Laptop	Stop Sign
Airplane	17 85.5%	2 10.0%	1 5.0%	0 0.0%
Ferry	3 15.0%	17 85.0%	0 0.0%	0 0.0%
Laptop	1 5.0%	0 0.0%	19 95.0%	0 0.0%
Stop Sign	0 0.0%	0 0.0%	1 5.0%	19 95.0%

Table 11: Confusion matrix for Fine KNN

	Airplane	Ferry	Laptop	Stop Sign
Airplane	12 60.0%	5 25.0%	2 10.0%	1 5.0%
Ferry	6 30.0%	11 55.0%	2 10.0%	1 5.0%
Laptop	1 5.0%	1 5.0%	9 45.0%	9 45.0%
Stop Sign	0 0.0%	1 5.0%	0 0.0%	19 95.0%

Table 12: Confusion matrix for Weighted KNN

	Airplane	Ferry	Laptop	Stop Sign
Airplane	19 95.0%	1 5.0%	0 0.0%	0 0.0%
Ferry	12 30.0%	7 55.0%	1 10.0%	0 0.0%
Laptop	10 55.0%	0 0.0%	9 45.0%	1 0.0%
Stop Sign	0 0.0%	1 5.0%	0 0.0%	19 95.0%

Table 13: Confusion matrix for Boosted Trees

	Airplane	Ferry	Laptop	Stop Sign
Airplane	8 40.0%	11 55.0%	1 5.0%	0 0.0%
Ferry	3 15.0%	14 70.0%	3 15.0%	0 0.0%
Laptop	1 5.0%	6 30.0%	11 55.0%	2 10.0%
Stop Sign	0 0.0%	2 10.0%	0 0.0%	18 90.0%

Table 14: Confusion matrix for Bagged Trees

	Airplane	Ferry	Laptop	Stop Sign
Airplane	16 80.0%	2 10.0%	2 10.0%	0 0.0%
Ferry	4 20.0%	16 80.0%	0 0.0%	0 0.0%
Laptop	0 0.0%	1 5.0%	18 90.0%	1 5.0%
Stop Sign	0 0.0%	1 5.0%	1 5.0%	18 90.0%

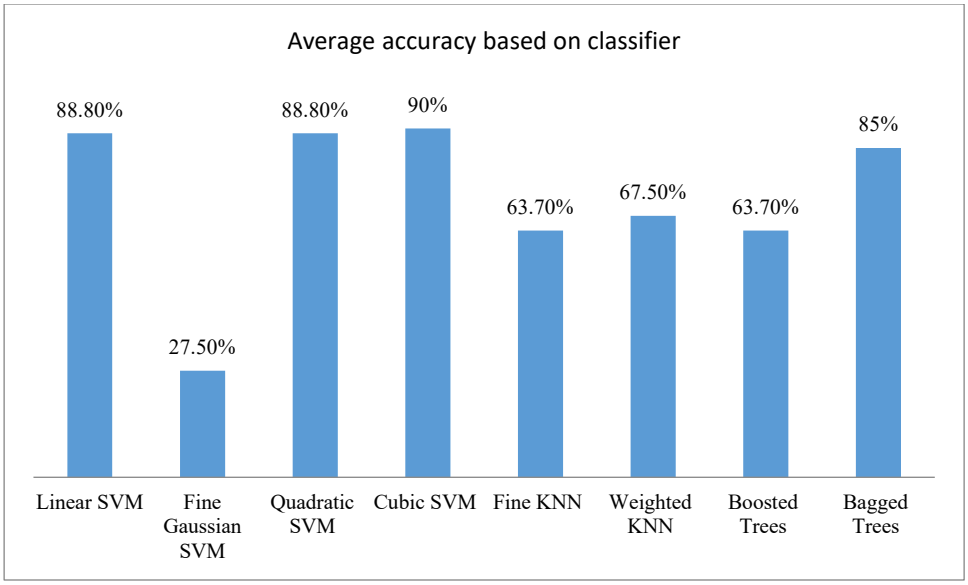


Figure 8: Average accuracy based on the learning classifier

7. Conclusion

In this paper, we related the different techniques and algorithms used in our machine learning framework for image classification. We presented machine learning state-of-the-art applied to image classification. We introduced the Bag of Features paradigm used for input image encoding and highlighted the SURF as its technique for image features extraction. Through experimentations we proofed that using SURF local feature extractor method for image vector representation and SVM (cubic SVM) training classifier performs best prediction average accuracy. In test scenarios we focused on stop sign image as we project to apply the trained classifier in a robotic system.

8. References

[1] L.Breiman, Statistical modeling: the two cultures. Statistical Science, 16, 199–215, 2001.

[2] A. Ramanathan, L. Pullum, H. Faraz, C. Dwaipayan, J. K. Sumit, "Integrating symbolic and statistical methods for testing intelligent systems: Applications to machine learning and computer vision" in Design, Automation & Test in Europe Conference & Exhibition (DATE), Dresden, Germany, 2016.

[3] S. Liu, X. Bai, Discriminative features for image classification and retrieval. Pattern Recognition. Lett. 33(6), 744–751, 2012.

[4] David Jardim; Luís Nunes; Miguel Dias, "Human Activity Recognition from automatically labeled data in RGB-D videos" in 8th Computer Science and Electronic Engineering (CEEC), Colchester, UK ,2016. https://doi.org/10.1109/CEEC.2016.7835894

[5] J. Stöttinger, A. Hanbury,N. Sebe, T. Gevers, "Sparse color interest points for image retrieval and object categorization" IEEE Trans. Image Process. 21(5), 2681–2691, 2012. https://doi.org/ 10.1109/TIP.2012.2186143

[6] J. Wang,Y. Li, Y. Zhang,C. Wang, H. Xie, G. Chen, X. Gao, Bag-of-features based medical image retrieval via multiple assignment and visual words weighting. IEEE Trans. Med. Imaging 30(11), 1996–2011, 2011. https://doi.org/10.1109/TMI.2011.2161673

[7] O. Miksik, K. Mikolajczyk, "Evaluation of local detectors and descriptors for fast feature matching" in International Conference on Pattern Recognition (ICPR 2012), pp. 2681–2684. Tsukuba, Japan, 2012.

[8] B. Kim, H. Yoo, K. Sohn, Exact order based feature descriptor for illumination robust image matching. Pattern Recognition. 46(12), 3268–3278, 2013.

[9] T T Dhivyaprabha, P. Subashini, M. Krishnaveni, "Computational intelligence based machine learning methods for rule-based reasoning in computer vision applications", in IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 2016. https://doi.org/10.1109/SSCI.2016.7850050

[10] P. Moreels, P. Perona, "Evaluation of features detectors and descriptors based on 3D objects", in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Beijing, China, 2005. https://doi.org/10.1109/ICCV.2005.89

[11] F.M. Campos, L. Correia, J.M.F. Calado, Robot visual localization through local feature fusion: an evaluation of multiple classifiers combination approaches. J. Intell. Rob. Syst. 77(2), 377–390, 2015.

[12] S. Zhang, Q. Tian, Q. Huang, W. Gao, Y. Rui, "USB: ultrashort binary descriptor for fast visual matching and retrieval",in IEEE Transactions on Image Processing, 2014. https://doi.org/ 10.1109/TIP.2014.2330794

[13] L. Fei-Fei ,P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories", in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) 2: 524. doi:10.1109/CVPR.2005.16. ISBN 0-7695-2372-2, 2005. https://doi.org/10.1109/CVPR.2005.16

[14] D. Jyothy, S. Martish, M. Leena, "Bag of feature approach for vehicle classification in heterogeneous traffic", in IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), Kollam, India, 2017. https://doi.org/10.1109/SPICES.2017.8091346

[15] S. Zhang; A. P. Leung, "A Novel approach to dictionary learning for the bag-of-features model", in International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Ningbo, China, 2017. https://doi.org/10.1109/ICWAPR.2017.8076672

[16] T. Lindeberg, Scale selection, Computer Vision: A Reference Guide, (K. Ikeuchi, ed.), Springer, pages 701-713, 2014.

[17] E. Rosten, T. Drummond, Machine learning for high-speed corner detection. European Conference on Computer Vision. Springer. pp. 430–443. https://doi.org/10.1007/11744023_34, 2006.

[18] N.Vidal-Naquet, Ullman, "Object recognition with informative features and linear classification" in Proceedings Ninth IEEE International Conference on Computer Vision (PDF),Nice, France, 2003. https://doi.org/ 10.1109/ICCV.2003.1238356

[19] R. Cuingnet, C. Rosso, M. Chupin, S. Lehéricy, D. Dormont, H. Benali, Y. Samson, O. Colliot, Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome, Medical Image Analysis, 15 (5): 729–737, 2011.

[20] D. Lowe, Towards a computational model for object recognition in IT cortex. Proc. Biologically Motivated Computer Vision, pages 2031, 2000.

[21] S. Mark, S. Nixon Alberto, Feature Extraction and Image Processing, Elsevier Ltd, ISBN: 978-0-12372-538-7, 2008.

[22] D.G Lowe, Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91–110, 2004.

[23] M.W David, Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation (PDF). Journal of Machine Learning Technologies 2 (1): 37–63, 2011.

[24] R. Piyush, Hyperplane based Classification: Perceptron and (Intro to) Support Vector Machines, CS5350/6350: Machine Learning September 8, 2011.

[25] C. Domeniconi, D. Gunopulos, J. Peng, “Large margin nearest neighbor classifiers” in IEEE Transactions on Neural Networks, 2005. https://doi.org/10.1109/TNN.2005.849821

[26] L. Rokach,O. Maimon, Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711, 2008.

[27] J.H Friedman, Greedy Function Approximation: A Gradient Boosting Machine, 1999.

[28] Z. Jinyi, L. Wei, C. Chen, D. Qian, Scene classification using local and global features with collaborative representation fusion; Information Sciences 348, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 11-17 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Impact of Line Resistance Combined with Device Variability on Resistive RAM Memories

Hassan Aziza*, Pierre Canet, Jeremy Postel-Pellerin

Aix Marseille Univ., CNRS, IM2NP UMR 7334, 60 rue F. Joliot-Curie, 13453 Marseille Cedex 13, France

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 13 December, 2017
Online: 18 January, 2018

Keywords:
Resistive RAM arrays
Variability
voltage drop (IR drop)

ABSTRACT

In this paper, the performance and reliability of oxide-based Resistive RAM (ReRAM) memory is investigated in a 28nm FDSOI technology versus interconnects resistivity combined with device variability. Indeed, common problems with ReRAM are related to high variability in operating conditions and low yield. At a cell level ReRAMs suffer from variability. At an array level, ReRAMs suffer from different voltage drops seen across the cells due to line resistances. Although research has taken steps to resolve these issues, variability combined with resistive paths remain an important characteristic for ReRAMs. In this context, a deeper understanding of the impact of these characteristics on ReRAM performances is needed to propose variability tolerant designs to ensure the robustness of the technology. The presented study addresses the memory cell, the memory word up to the memory matrix.

1. Introduction

Because data storage and processing solutions are so central to modern technology, many research works are dedicated to pursue new types of computer memory. One of the major goal is to develop a universal memory (i.e. a storage medium that would combine the high speed of RAM with the non volatility of a Flash drive). Additionally, embedded memory is a fundamental component of any electronic system including high-performance System-on-Chip (SoC) and Internet of Things (IoT) devices. In this context, the presented work proposes a reliability analysis of the Resistive RAM memory which considered as a potential universal memory candidate. This study is an extension of work originally presented in [1], with an extension to a whole memory matrix. Moreover, a multilevel storage capability of ReRAM cells is demonstrated at a matrix level.

According to ITRS [2], embedded Non Volatile Memories (NVMs) are occupying a major part of the area of a typical (SoC). Although Flash memory is widely used today, it needs high voltage for Write and Erase operations and has reliability issues that are hard to handle, increasing the cost of circuit design and process integration. Thus, the industry is trying to find a good alternative NVM that can replace Flash memories. Possible candidates include Magnetic RAM (MRAM), ReRAM, Phase Change Memory (PCM), Ferroelectric RAM (FeRAM), etc. Compared to MRAM and PCM technologies, ReRAM technology is still in an emerging phase [3][4]. However, a considerable technological effort is currently driven worldwide to push this technology to prototype level. ReRAM advantages comprise non volatile data storage at low power and latency and high memory density while maintaining device performance and reliability [5]. Moreover, the 3D-staking technology developed for Flash memories can be transferred to ReRAM and the multilevel cell operation scheme of Flash memories can also be achieved in ReRAMs [6].

However, the continuous push for scalability to obtain high-density chips makes the ReRAM technology extremely sensitive to variability, physical defects and environmental influences that may severely compromise its correct behavior [7]. At sub 32nm node, size reduction increases the resistivity of interconnects, inducing a voltage drop along the memory matrix lines, which can cause reliability issues. Indeed, as ReRAM data is stored as two resistance states of the resistive switching device, these memories are sensitive to resistive paths [8]. On top of that, common problems with ReRAMs are related to high variability in operating conditions [9].

In memory devices relying on resistance change such as ReRAMs, complex physical mechanisms are responsible for reversible switching of the electrical conductivity between high and low resistance states. This resistivity change is generally attributed to the formation/dissolution of conductive paths between metallic electrodes [10]. A typical ReRAM device consists of two metallic electrodes that sandwich a thin dielectric layer serving as permanent storage medium making its leakage current close to

*Corresponding Author: Hassan Aziza, Email: hassen.aziza@univ-amu.fr

zero. Oxide-based Resistive Random Access Memory (so-called OxRAM) use transition metal oxides as a dielectric layer. In this study, an HfO_2 Oxide-based ReRAM stack is considered [11].

OxRAM cell operation is depicted in Figure 1. After an initial electroforming process, the memory element may be reversibly switched between two distinct resistance states. Electroforming stage corresponds to a voltage-induced resistance switching from an initial very high resistance state (virgin state) to a conductive state. After FORMING, resistive switching corresponds to an abrupt change between a High Resistance State (HRS or OFF state) and a Low Resistance State (LRS or ON state). This resistance change is achieved by applying specific voltage (i.e. V_{SET} and V_{RES}) to SET and RESET the memory cell.

It is important to note that the FORMING stage is the first and most critical step as it determines the switching characteristics during the future operation of the memory cell. Thus, the Forming Resistance State (R_{FRS}) which characterizes the filament creation is a key parameter in terms of reliability. Besides, the forming step requires high voltage levels (more important than V_{SET} and V_{RES}).

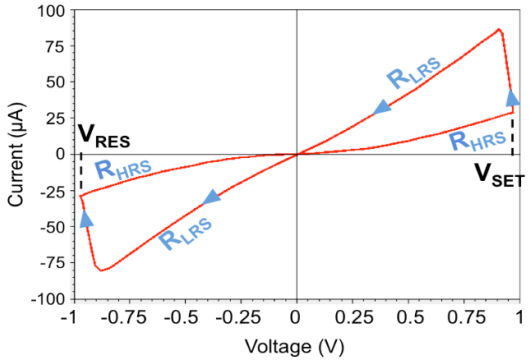


Figure 1. Typical I-V characteristic of a bipolar OxRAM cell

In this paper, the reliability of a ReRAM memory array is investigated versus interconnects and device variability. Section II presents the OxRAM model used for simulations. Section III address a 32-bit memory word. In section III, the study is extended to the memory matrix and the multilevel storage capability of ReRAM cells is demonstrated. Section IV concludes the paper. For each section simulation results based on a 28nm Fully-Depleted Silicon-On-Insulator (FDSOI) are proposed.

2. ReRAM elementary cell model

The proposed OxRAM modeling approach relies on electric field-induced creation/destruction of oxygen vacancies within the switching layer. The model enables continuously accounting for both SET and RESET operations into a single master equation in which the resistance is controlled by the radius of the conduction pathways (r_{CF}) [12]:

$$\frac{dr_{CF}}{dt} = \left(r_{CF\max} - r_{CF}\right)\cdot 10^{\beta_{RedOx}}\cdot e^{\frac{Ea - q\cdot\alpha_{red}\cdot V_{cell}}{k_b\cdot T}} - r_{CF}\cdot 10^{\beta_{RedOx}}\cdot e^{\frac{Ea + q\cdot\alpha_{ox}\cdot V_{cell}}{k_b\cdot T}} \quad (1)$$

Where β_{RedOx} is the nominal oxide reduction rate, E_a is the activation energy, α_{red} and α_{ox} are the transfer coefficients (ranging between 0 and 1 and representing the pathways creation/destruction dynamic), k_b is the Boltzmann constant, r_{CFmax} is the maximal size of the conductive filament radius, T is the temperature and V_{cell} the voltage across the cell.

Moreover, the model makes assumptions of a uniform radius of the conduction pathways, a uniform electric field in the cell and temperature triggered acceleration of the oxide reduction reactions ("redox"). Finally, the total current in the OxRAM includes two components, i.e. one is related to the conductive species (I_{CF}) and the other to the conduction through the oxide (I_{OX}).

$$I_{CF} = \frac{V_{Cell}}{L_x}\cdot\left(\pi\cdot r_{CF}^2\cdot\left(\sigma_{CF} - \sigma_{OX}\right) + \pi\cdot r_{CF\max}^2\cdot\sigma_{OX}\right) \quad (2)$$

$$I_{OX} = A_{HRS}\cdot S_{Cell}\left(\frac{V_{Cell}}{L_x}\right)^{\beta_{HRS}} \quad (3)$$

where L_x is the oxide thickness, S_{Cell} is the total area of the device, σ_{Ox} the oxidation rate and σ_{CF} the reduction rate. To take into account I_{OX} trap assisted current (Poole-Frenkel, Schottky emission, Space Charge Limited Current (SCLC)), a power law between the cell current and the applied bias is considered with two parameters A_{HRS} and β_{HRS}. Finally, the total current flowing through the cell is:

$$I_{Cell} = I_{CF} + I_{OX} \quad (4)$$

I_{CF} is the main contributor to LRS current (I_{LRS}) and I_{OX} is the main contributor to HRS current (I_{HRS}).

The memory cell compact model is calibrated on silicon. The model was confronted to quasi-static and dynamic experimental data before its implementation in electrical circuit simulators. As presented in Figure 2a (current voltage characteristic in logarithmic scale), after calibration, the model satisfactorily matches quasi-static and dynamic experimental data measured on actual HfO_2-based memory elements (TiN/Ti/HfO_x/TiN stack). In Figure 2b, the evolution of SET voltages (V_{app}) as a function of the programming ramp speed is presented. The model implementation focused on this dependence which is crucial for the model to be confidently implemented in circuit simulators.

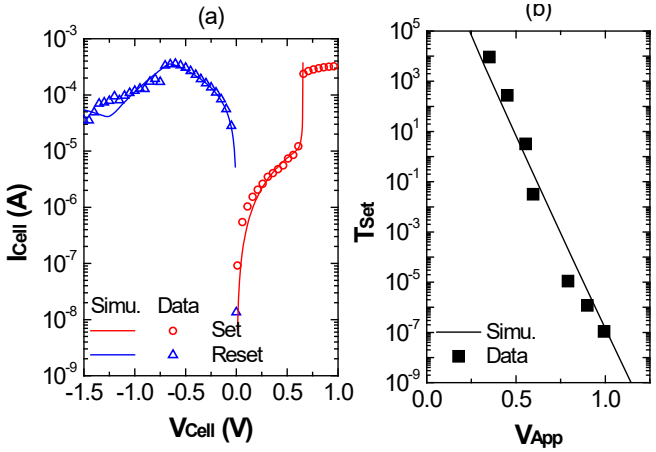


Figure 2. (a) I-V characteristic measured on HfO2-based devices and corresponding simulation using a bipolar OxRAM physical model. (b) SET voltage as a function of the programming ramp

Due to the stochastic nature of the switching process in OxRAMs, leading to large variability, the OxRAM model features a variability dependency. The variation is chosen to fit experimental data as presented in Figure 3. The model behaviour (lines) is consistent with experimental data (symbols). The cell variability modelling is based on OxRAM card model parameters variation. Variability is introduced through specific model parameters (*i.e.* β_{RedOx} and β_{HRS} parameters presented respectively in Equation 1 and Equation 3). Moreover, at sub 32nm node, MOSFET mismatch in the transistor subsystem (digital and analog blocks) increases inherent variability of OxRAM circuits, increasing the overall variability.

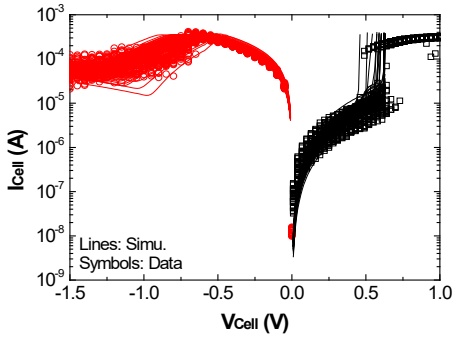


Figure 3. Measured and corresponding simulated I-V characteristic obtained from TiN/Ti/HfO2/TiN devices showing strong variation on R_{LHS} and R_{HRS}

3. ReRAM memory word

3.1. Memory word architecture

A 1T1R ReRAM cell (one MOS Transistor in series with one Resistor) placed in a 32-bit word is considered and presented in Figure 4. The word selection is achieved through the Word Line (WL) before any operation. Word programming is performed in 2 steps, considering that the FORMING operation occurs only once in the product lifetime. Once the word is selected, all cells are RESET in parallel (logical "0") through the Reset Word Line (WL_R), then memory cells are all SET (logical "1") through the Bit Line (BL). The WL_R line drives the whole word RESET current, making it sensitive to line resistivity.

Indeed, during a memory word program operation, a voltage drop occurs along the WL_R line, which can be critical in terms of programming efficiency. To monitor the programming efficiency, High Resistance State (R_{HRS}) and Low Resistance State (R_{LRS}) resistances are extracted after RESET and SET operations respectively.

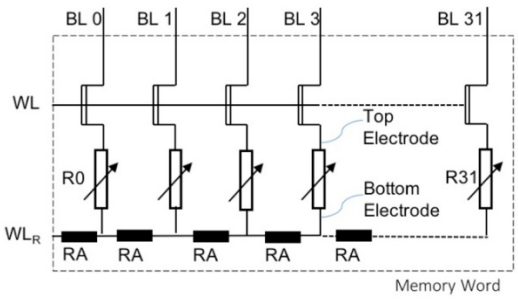


Figure 4. 32-bit ReRAM memory word

The memory word is simulated using the OxRAM model presented in section 2. The model is calibrated on silicon for a 28nm FDSOI technology. With respect to the programming conditions, typical values for LRS and HRS are R_{LRS}=10kΩ and R_{HRS}=130kΩ for the considered technology. The WL_R line resistance per cell (including vias) is called R_A (see Figure 4) and is evaluated to 1.6Ω [8].

3.2. Simulation results

Impact of line resistances on R_{HRS} and R_{LRS} is presented in Figure 5. The voltage drop induced by line resistances results in a weak FORMING (i.e. high resistive path between ReRAM electrodes). SET and RESET programming levels are thus impacted resulting in a higher R_{HRS} (+5%) and R_{LRS} (+30%) values for cells located at the end of the memory word.

Impact of line resistances combined with variability is presented in Figure 6. After 300 Monte Carlo runs, R_{HRS} and R_{LRS} distributions are extracted for 1 cell over 4 and displayed in box-plot forms. Figure 6a shows that LRS distribution spread increases along the word line. Besides, anomalous LRS residual cell populations (far from their typical LRS values, represented by dashed lines) are visible from cell 7. This effect is related to cells not properly formed due to line resistances. In Figure6b, HRS distributions are presented. One can notice the HRS mean distribution shift to higher values with the word length.

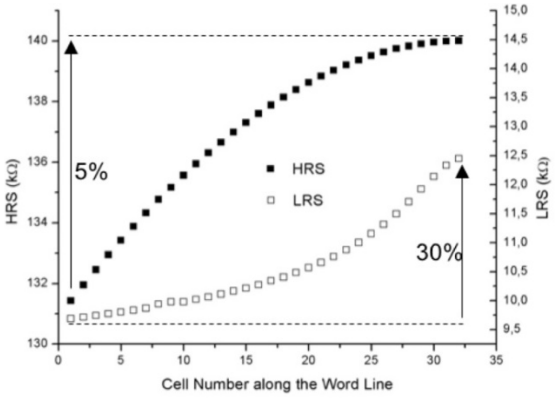


Figure 5. Resistances (HRS & LRS) along the WL

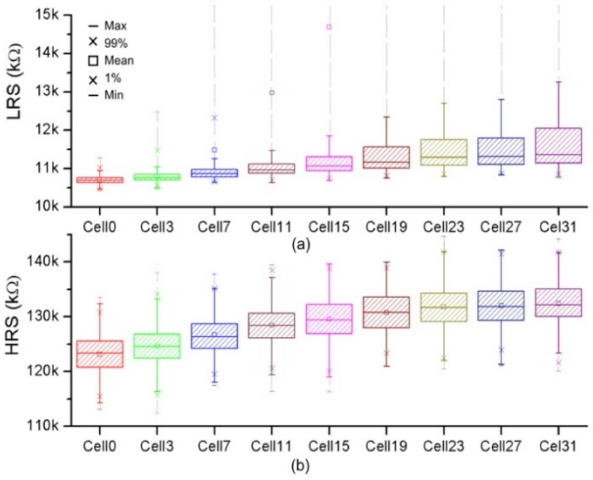


Figure 6. (a) LRS and (b) HRS distributions along the WL

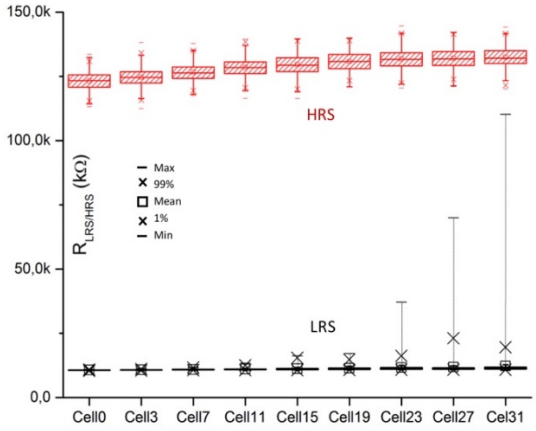


Figure 7. LRS and HRS distributions overlaps along the WL

For comparison purposes, Figure 7 presents the full range variation (from min to max) of each box-plot (HRS and LRS states). Along the WL, HRS/LRS distributions are getting closer and closer due to marginal LRS cells, affecting the memory reliability (i.e. HRS distribution generally larger than LRS distribution is much more degraded by resistive paths).

4. ReRAM memory array

4.1. Memory array architecture

Figure 8 presents the elementary array used for simulation which is constituted by a 3×3 1T1R cell matrix, a row decoder, a column decoder and a sense amplifier for the read operation. The memory cell is modeled by the OxRAM cell model presented in section 2.

The variability analysis is conducted through Monte Carlo simulations. Each Monte Carlo simulation targets specific elements of the circuit: the resistive element, the select transistor, the decoding blocs and the sense amplifier. As a result, R_{ON} and R_{OFF} distribution spreads are extracted. Thus, the contribution of each element in terms of memory performance degradation is demonstrated. A 28-nm Fully Depleted Silicon-On-Insulator (FDSOI) technology is considered for simulations [13].

Memory array cells are first placed in a virgin state. Then, the memory array programming is done in 2 cycles. First, all memory cells are set (logical "1"), then the memory array is reset (logical "0"). R_{ON} value is extracted after the SET operation and R_{OFF} after the RESET operation for each Monte Carlo run.

Variability introduced in the resistive element is chosen to feet experimental date whereas variability of the transistor subsystem is given by the considered technology.

4.2. Simulation results

In this study, 4 different configurations are considered: the "ALL" configuration considers variability in the memory cell (resistive element and select transistor) and the peripheral circuits. In the "CELL" configuration, variability affects only the resistive element. In the "SELECT" configuration, variability affects only the OxRAM select transistor. In the "NO CELL" configuration, variability affects only the peripheral circuits (transistor subsystem, including the select transistor). In this study, the impact of the select transistor variability is subject to special attention. Indeed, the select transistor compliance allows the control of the maximum available current during the set transition, impacting directly the ON/OFF resistance. Table 1 summarizes the 4 configurations to simulate.

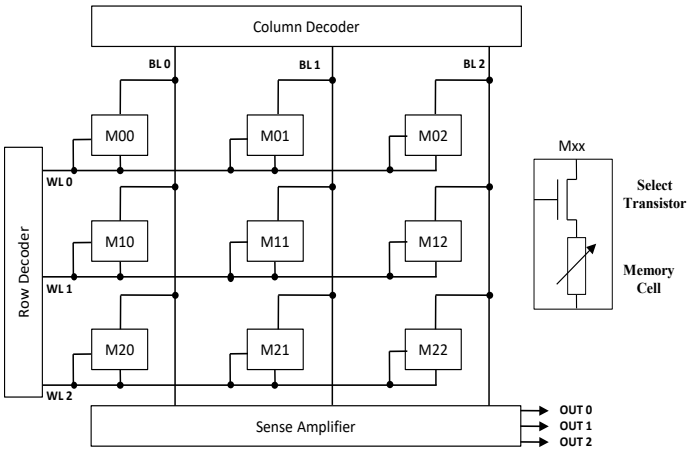


Figure 8. 3×3 OxRAM elementary memory array

Table 1: Simulated Configurations

Configuration	RUNS	Elements under variability
« ALL »	400	All circuit elements
« CELL »	400	Resistive element
« SELECT »	400	Select transistor
« NO CELL »	400	All elements excluding the resistive element

Figure 9 presents the impact of variability of the whole circuit elements (resistive element and transistor subsystem: "ALL" configuration) on R_{ON} and R_{OFF} distributions. Note that R_{OFF} distribution is much larger than R_{ON} distribution. In order to discriminate the contribution of each element of the circuit on ON/OFF resistances, a set of Monte Carlo simulations are also performed with "CELL", "SELECT" and "NO CELL" configurations.

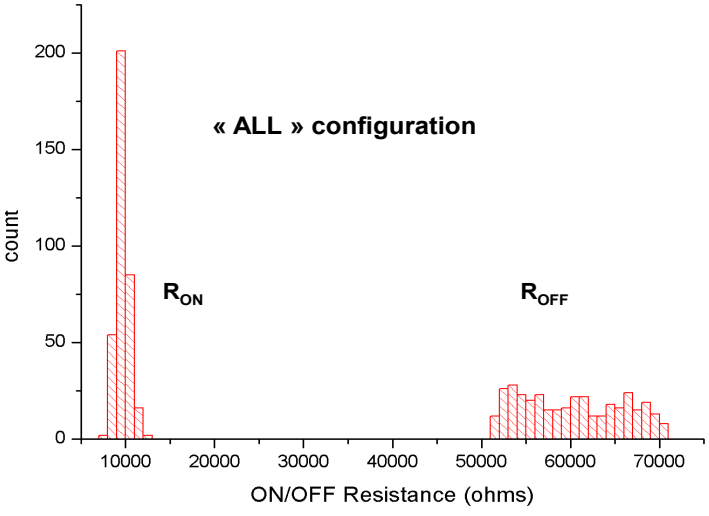


Figure 9. ON/OFF resistance distributions for the "ALL" configuration after 400 Monte Carlo simulations.

Figure 10 presents the impact of variability on R_{OFF} distributions (after RESET) for the 4 configurations. By comparing configuration "ALL" with configuration "CELL", it appears clearly that the impact of cell variability is more important than the impact of the transistor subsystem variability. The third and fourth configurations ("NO CELL" and "SELECT") shows that impact of transistors variability on R_{OFF} is negligible. Besides, the contribution of the select transistor variability is dominant in the peripheral circuit elements. Table 2 proposes a synthesis of results obtained in Figure 10 (mean values, standard deviation and standard deviation reported to the maximum standard deviation value).

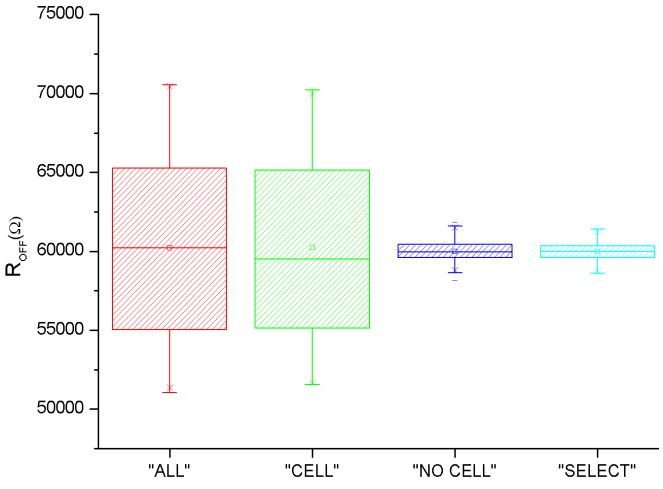


Figure 10. R_{OFF} distributions after 400 Monte Carlo simulations.

Table 2: R_{off} Distribution parameters

Configuration	Mean (Ω)	σ (Ω)	σ (%)
« ALL »	60200	5598	100
« CELL »	60240	5617	99
« NO CELL »	60004	598	9.6
« SELECT »	59990	536	10.6

Figure 11 presents the impact of variability on R_{ON} distributions (after SET) for the 4 configurations.

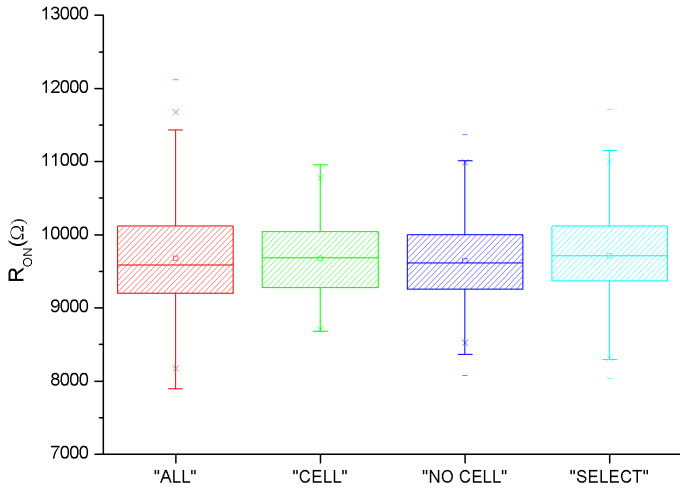


Figure 11. R_{ON} distributions after 400 Monte Carlo simulations.

Table 3: R_{on} Distribution Parameters

Configuration	Mean (Ω)	σ (Ω)	σ (%)
« ALL »	9674	724	100
« CELL »	9679	513	70.8
« NO CELL »	9642	550	78.7
« SELECT »	9713	570	75.9

Compared to Figure 10, it appears clearly that the cell variability impact is less pronounced for low resistive states (LRS), which is a common feature of all OxRAM technologies [14-15]. Moreover, the contribution of transistors variability is comparable to the contribution of the memory cell variability (see "CELL" and "NO CELL" configurations in Figure 11). Here again, simulation results show that the impact of the select transistor is not negligible. Table 3 proposes a synthesis of results obtained in Figure 11.

Figure12 shows a comparison between R_{ON} and R_{OFF} distributions. Based on the considered technology, it is shown that resistance variability in RESET state is much more important than variability in the SET state. This variability is mainly due to the memory cell. In addition, the impact of the select transistor is non-negligible. This effect is visible in the SET state where the impact of the memory cell variability is much less important, making the select transistor variability critical.

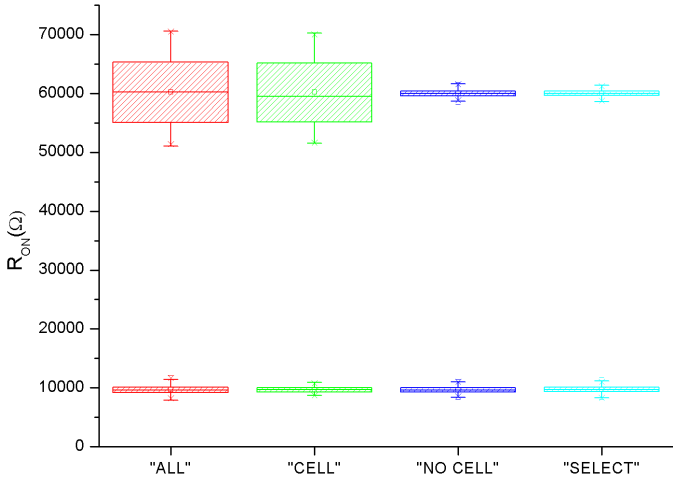


Figure 12. R_{ON}/R_{OFF} distributions after 400 Monte Carlo simulations.

4.3. Multi level approach

When a voltage is applied across a ReRAM cell, depending upon the voltage polarity, one or more Conductive Filaments (CFs) made out of oxygen vacancies are either formed or ruptured. Once the CFs are formed inside the metal oxide to bridge the top and bottom electrodes, current can flow through the CFs, and the cell is in a low resistance state. The larger the size of the CFs, the lower the resistance. Conversely, the rupture of the CFs disconnects the top electrode from the bottom electrode, resulting in a high resistance state (HRS) of the cell [16].

According to Equation 1, CF size is directly linked to the voltage across the cell, thus, multilevel LRS can be achieved in the 1T1R RRAM by modulating the SET voltage. In this study, the amplitude of the RESET voltage remains unchanged during the

cell programming operation. As HRS variation is generally larger than LRS variation, multilevel HRS is not considered. The HRS state is associated with an initial value, restored before each SET operation. Finally, to obtain intermediate SET states, programming is done in 2 steps: a RESET operation is first performed to switch the memory cell in its HRS initial state, then a SET operation sets the desired LRS level.

At a circuit level, different implementations of Multi-Level Cell (MLC) operation can be adopted. Multilevel operation can be achieved:

- By applying an increasing number of identical voltage pulses through the SET decoder. In this case, LRS level is function of the number of pulses [17].
- By modulating the gate voltage (WL) of the memory cell Select Transistor to control the SET current [18].
- By modulating directly the SET voltage generated from the SET decoder. In this case, LRS levels are achieved with different SET voltage values.

The first approach requires a pulse generator circuitry while the two others require different voltage levels generated from a specific circuitry. In this study, the third approach is considered as it is the most effective MLC approach for the considered technology. Four different LRS levels are generated by changing the amplitude of the SET voltage. Table 4 presents the different SET and RESET voltages values with the corresponding resistance nominal values.

Table 4: Programming Voltages and corresponding Resistance Values

	SET				RESET
Logical value	**00**	**01**	**10**	**11**	-
$V_{SET/RESET}$ (V)	1.4	1.8	2.2	2.8	-2.8
$R_{ON/OFF}$ (kΩ)	104	77	46	30	196

To assess the robustness of the analog resistance values, a Monte Carlo (MC) analysis is conducted using a 28-nm Fully Depleted Silicon-On-Insulator (FDSOI) technology. As a result, R_{ON} and R_{OFF} distribution spreads are extracted. The programming protocol is comparable to the one used in section 4.2. The memory array programming is done in 2 cycles. First, all memory cells are RESET (with an more important RESET voltage level), then the memory array is SET with specific SET voltage levels.

Figure 13 presents the impact of the memory cell variability on R_{ON} and R_{OFF} distributions after 300 MC simulations. Note that R_{OFF} distribution is much larger than R_{ON} distribution, preventing a HRS MLC implementation. R_{ON} distributions are narrower with a tendency to spread for small SET voltage values. In order to discriminate the different memory states during the read operation, no distribution overlap is permitted.

Figure 14 focuses on R_{ON} distributions and shows distinct R_{ON} resistance levels. Distributions are displayed in box-plot forms. The full range variation (from min to max) of each box-plot confirms the clear separation between each resistance level.

Figure 13 and Figure 14 results are presented for cell variability parameters included in the range ±15% of the standard deviation of a normal distribution (i.e. $\sigma_{Variability}$ = 15%). This variability is consistent with silicon data (see Figure 3) which means that a MLC approach can be considered for this technology.

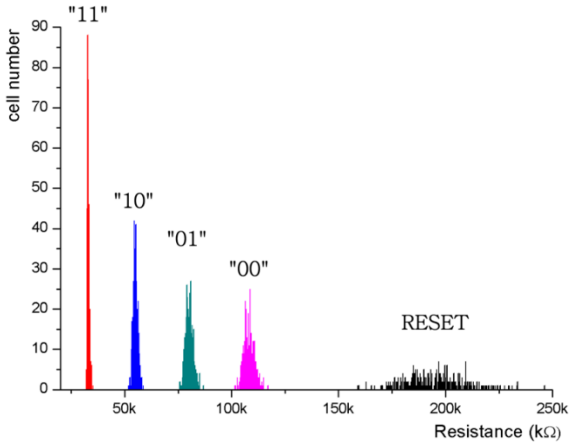


Figure 13. R_{ON} and R_{OFF} distributions versus cell variability ($\sigma_{Variability}$ = 15%) after 300 runs (SET and RESET states)

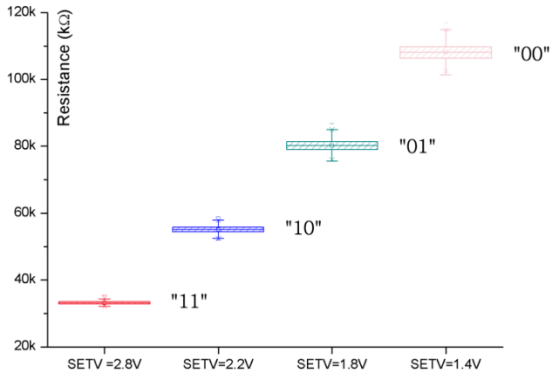


Figure 14. R_{ON} distribution box-plots versus cell variability ($\sigma_{Variability}$ = 15%) after 300 runs (SET states)

To evaluate the robustness of the technology, Figure 15 and Figure 16 results are provided for $\sigma_{Variability}$ = 20% and $\sigma_{Variability}$ = 25%, which is the worst case. An important spreading of R_{ON} is observed. The initial R_{ON} distribution spread increases as variability increase. Even for the worst case ($\sigma_{Variability}$ = 25%), no overlap between LRS distributions is observed which confirms the robustness of MLC approach. In Table 5, R_{ON} (SET) and R_{OFF} (RESET) distribution parameters (mean value and standard deviation) are reported for the different variability parameter values.

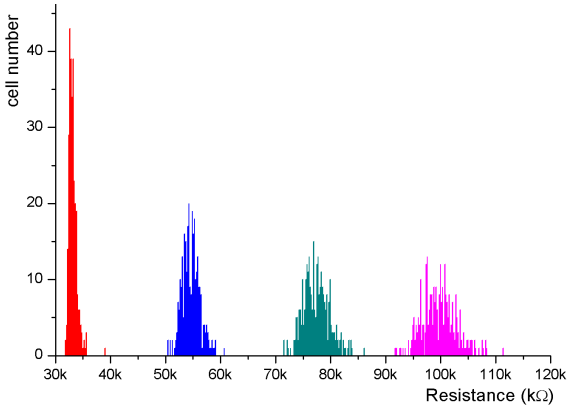


Figure 15. R_{ON} distributions versus cell variability ($\sigma_{Variability}$ = 20%) after 300 runs (SET states)

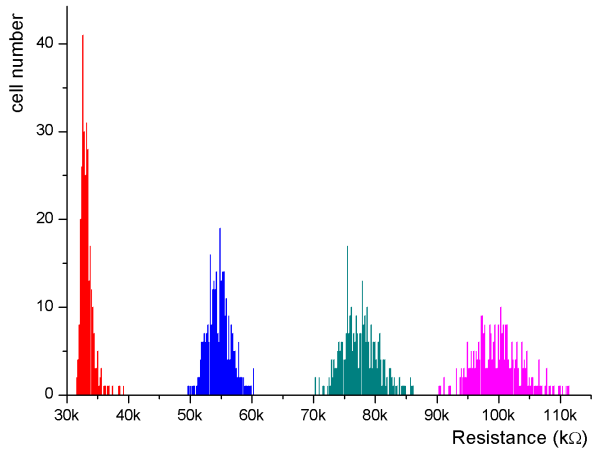


Figure 16. R_{ON} distributions versus cell variability ($\sigma_{Variability}$ = 25%) after 300 runs (SET states)

Table 5: Simulated Configurations

		SET				RESET
$\sigma_{Variability}$	Distributions	**11**	**10**	**01**	**00**	
15%	Mean (kΩ)	33.159	55.111	80.227	108.119	195.350
	σ (kΩ)	0.501	1.153	1.8550	2.5520	14.447
20%	Mean (kΩ)	33.239	54.681	77.402	99.672	196.639
	σ (kΩ)	0.768	1.614	2.408	3.174	19.520
25%	Mean (kΩ)	33.342	54.761	77.529	99.844	198.031
	σ (kΩ)	1.049	1.973	2.978	3.928	24.882

5. Conclusion

Variability combined with voltage drops due to resistive paths present a major challenge for ReRAM memory fabrication process and design engineers. In the proposed study, impact of variability combined with resistive paths is evaluated based on a 1T1R ReRAM elementary memory array. A specific attention is provided to the ReRAM memory word. To assess the technology reliability, the HRS/LRS resistance ratio is extracted at a memory word and at a memory array level to provide a valuable feedback for designers during ReRAM memory array design. At a word level, impact of line resistances combined with variability can affect the memory word operation from a word size greater than 8 bits. At a memory array level, resistance variability in RESET state is much more important than variability in the SET state. Additionally, a stable MLC operation of 2-bits/cell is demonstrated.

References

[1] H. Aziza, P. Canet, J. Postel-Pellerin, M. Moreau, JM. Portal, M. Bocquet, "ReRAM ON/OFF resistance ratio degradation due to line resistance combined with device variability in 28nm FDSOI technology", Ultimate Integration on Silicon (EUROSOI-ULIS), 2017. DOI: 10.1109/ULIS.2017.7962594.

[2] Semiconductor Industry Association, "International Technology Roadmap for Semiconductors".

[3] E. Shahrabi, B. Attarimashalkoubeh, J. Sandrini, and Y. Leblebici, "Towards chip-level reram-cmos co-integration", in International Conference on Memristive Materials, Devices and Systems (MEMRYSIS), 2017. DOI: 10.1109/PRIME.2016.7519497.

[4] H. Aziza, H. Ayari, S. Onkaraiah, M. Moreau, JM. Portal, "Multilevel operation in oxide based resistive RAM with SET voltage modulation", in International Conference on Design and Technology of Integrated Systems in nanoscale Era, 2016. DOI: 10.1109/DTIS.2016.7483892.

[5] Waser, R., Nanoelectronics and Information Technology: John Wiley & Sons, 2012.

[6] Chen, F., Seok, J.Y. and Hwang, C.S., Integration Technology and Cell Design, in Resistive Switching, Wiley-VCH Verlag GmbH & Co. KGaA, 573-596, 2016. DOI: 10.1002/9783527680870.ch20.

[7] Molas, G., et al. "Functionality and reliability of resistive RAM (RRAM) for non-volatile memory applications." 2016 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA). IEEE, 2016. DOI: 10.1109/VLSI-TSA.2016.7480520.

[8] Liang & al., "Effect of wordline/bitline scaling on the performance, energy consumption, and reliability of cross-point memory array", ACM Journal on Emerging Technologies in Computing Systems, Vol. 9, No. 1, 2013. DOI: 10.1145/2422094.2422103.

[9] E.I. Vatajelu et al., "Nonvolatile memories: Present and future challenges", Design & Test Symposium (IDT), 61-66, 2014. DOI: 10.1109/IDT.2014.7038588.

[10] Hyo-Shin Ahn1 et al., "Pairing of cation vacancies and gap-state creation in TiO_2 and HfO_2," Appl. Phys. Lett., Vol. 90, Iss. 25, p. 252908 - 252908-3 (2007). DOI: 10.1063/1.2749858.

[11] C. Lien et al., "The highly scalable and reliable hafnium oxide ReRAM and its future challenges", Solid-State and Integrated Circuit Technology (ICSICT), 2010. DOI: 10.1109/ICSICT.2010.5667553.

[12] M. Bocquet et al., "Compact Modeling Solutions for Oxide-Based Resistive Switching Memories (OxRAM)", Journal of Low Power Electronics and Applications, 4 (1), 1-14, 2014. DOI: 10.1109/FTFC.2013.6577779.

[13] C. Mazure, et al., "FDSOI: From substrate to devices and circuit applications", ESSCIRC, 45-51, 2010. DOI: 10.1109/ESSCIRC.2010.5619767.

[14] F. Nardi, et al, "Resistive switching by voltage-driven ion migration in bipolar RRAM Part I: Experimental study", IEEE Transactions on Electron Devices, 59(9), 2461-2467, 2012. DOI: 10.1109/TED.2012.2202319.

[15] S. Larentis, et al., "Resistive switching by voltage-driven ion migration in bipolar RRAM Part II: Modeling", IEEE Trans. on Electron Devices, 59(9), 2468-2475, 2012. DOI: 10.1109/TED.2012.2202320.

[16] T. Diokh, "Investigation of the Impact of the Oxide Thickness and RESET conditions on Disturb in HfO2-RRAM integrated in a 65nm CMOS Technology" in International Reliability Physics Symposium, 3-6, 2013. DOI: 10.1109/IRPS.2013.6532043.

[17] Y.C. Huang et al., "Using binary resistors to achieve multilevel resistive switching in multilayer NiO/Pt nanowire arrays", NPG Asia Materials, 6 (2), e85, 2014. DOI:10.1038/am.2013.81.

[18] Y. S. Fan et al., "High Endurance and Multilevel Operation in Oxide Semiconductor-Based Resistive RAM Using Thin-Film Transistor as a Selector", ECS Solid State Letters, 4(9), Q41-Q43, 2015. DOI: 10.1149/2.0061508ss.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 18-24 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Building an Efficient Alert Management Model for Intrusion Detection Systems

El mostapha Chakir[*,1], Mohamed Moughit[2,3], Youness Idrissi Khamlichi[4]

[1]*IR2M Laboratory, FST, Univ Hassan 1, Settat, Morocco*

[2]*IR2M Laboratory, ENSA, Univ Hassan 1, Settat, Morocco*

[3]*EEA&TI Laboratory, FST, Univ Hassan 2, Mohammedia, Morocco*

[4]*LERS Laboratory, ENSA, Univ Sidi Mohamed Ben Abdellah, FES, Morocco*

ARTICLE INFO

Article history:
Received: 14 November, 2017
Accepted: 13 December, 2017
Online: 18 January, 2018

Keywords:
Intrusion detection system,
False positive
Risk Assessment
Alerts prioritization
KDD'99
Pattern matching

ABSTRACT

This paper is an extension of work originally presented in WITS-2017 CONF. We extend our previous works by improving the Risk calculation formula, and risk assessment of an alert cluster instead of every single alert. Also, we presented the initial results of the implementation of our model based on risk assessment and alerts prioritization. The idea focuses on a new approach to estimate the risk of each alert and a cluster of alerts. This approach uses indicators such as priority, reliability and asset value as decision factors to calculate alert's risk. The objective is to determine the impact of alerts generated by Intrusion detection system (IDS) on the security status of an information system, and also improve the detection of intrusions using snort IDS by classifying the most critical alerts by their levels of risk. Thus, only alerts that present a real threat will be displayed to the security administrator. The implementation of this approach will reduce the number of false alerts and improve the performance of the IDS.

1. Introduction

This paper is an extension of work originally presented in International Conference on Wireless Technologies, Embedded and Intelligent Systems, EEE WITS-2017 [1]. Based on our work in [2], the goal is to improve the intrusion detection system (IDS) with a Risk Assessment method that can help to prioritize the generated alerts by their importance.

IDS systems generate enormous number of alerts. Often, there are duplicative events from various systems, and other alerts that could be characterized as noise (False Positive). A False Positive is normal events being classified as attacks. This is a major problem for many organizations [3,4]. An attack may in fact be happening, and the network administrator needs to be able to properly identify it, thus, he need to analyze each IDS alert manually, whether it is a false or true positive. So, it is a quite time consuming. Since the number of false positives is high, so alerts of real attacks are hidden among them. The optimal way to deal with this problem is to use an IDS solution that has the ability to prioritize alerts, calculate the risk of each one and correlate them, thereby to help the network administrator focus the efforts on detecting actual threats [5-7]. Therefore, an automated operation is needed to handle alerts and give a good perdition to the security administrator.

In this work, we propose the new Risk Assessment model as a method of prioritizing alerts according to the risk level of each one, and also evaluate the risk of a cluster of alerts. The risk is evaluated as a combination of certain parameters extracted from each alert.

This paper is organized as follows: Section 2 discusses the related work in risk assessment and alert prioritization; Section 3 presents the proposed model and the indicators that are used to support it; Section 4 discusses the implementation of the model and the experimental results to validate the proposed approach; and finally, in Section 5, we summarize the conclusions derived from this work and indicate possible future works.

2. Related Works

Intrusion detection system has an important role in the security and perseverance of active defense system against intruder attacks. They evolved from packet sniffers, used as a network troubleshooting tool to locate malfunctioning equipment and

[*]Corresponding Author: El Mostapha Chakir, IR2M Laboratory, FST, Univ Hassan 1, Settat, Morocco | Email: e.chakir@uhp.ac.ma

software by creating logs showing the activity of network packets [8,9]. Prior to the advent of network switches, IDS products could be connected to any port on a network hub and had a good chance of monitoring network packets on a local area network segment. Many researchers have proposed and implemented various models for IDS, but they often generate too many false positives due to their simplistic analysis [10].

Attacks are presented to a security administrator as alerts generated by IDSs. An IDS generates a large number of alerts and with this large number, security administrators are overwhelmed and it becomes difficult to manually distinguish between the real attacks and the false ones. To deal with this problem, two solutions have been proposed. The first one focuses on the monitoring device by enhancing its detection mechanism, optimizing its signatures, and choosing the right location [11]. Although this solution promises to reduce the number of alerts, it requires prior security administrator knowledge of detection mechanism. The second solution focuses on the sensor's outputs. Several IDS alert management techniques fall into this category and include aggregation, clustering, correlation and fusion [12].

Generally, reducing the number of false positives and prioritizing the most critical attacks are the main objectives of IDS alert management approaches. Furthermore, these techniques help the security administrators in understanding the situation revealed by the IDS.

In our previous work [2] we presented a new model to handle Intrusion detection system alerts based on a stateful pattern matching algorithm. In this paper we improve that model by proposing a new method based on risk assessment and alert prioritization using parameters extracted from alerts generated by IDS.

Risk assessment is the process of identifying security risks related to a system and determining their probability of occurrence, their impact, and the safeguards that would mitigate that impact [13]. Risks can be defined as the potential that a given threat will exploit vulnerabilities of an asset or group of assets to cause loss or damage to the assets. The main objective of risk assessment is to define appropriate controls for reducing or eliminating those risks.

Researchers have been proposed many approaches to prioritize alerts. In [13], the authors proposed a model that estimates the risk index for every single alert based upon indicators and input obtained from asset environments and attributes within the alerts themselves. The model uses a combination of two decision factors: impact on asset and likelihood of threat and vulnerability.

In [14], the authors proposed a method that evaluates IDS alerts based on a number of criteria. They used a Fuzzy Logic Inference mechanism in order to score alerts and developed a rescoring technique that enabled them to rescore alerts to show the early steps of the attackers, and applied their approach to the alerts generated by scanning DARPA 2000 LLDOS 1.0 dataset and they successfully prioritized the most critical alerts along with their preparation steps.

In [15], the authors applied a fuzzy system approach in accessing relative potential risks by associating potential vulnerabilities like the National Vulnerability Database (NVD) and Common Vulnerabilities and Exposures (CVE) with computer network assets.

In [16], the authors estimated risks by associating three criteria: computer network assets, attacks and vulnerabilities.

In [17], the authors proposed a model called "M-Correlator", an alert ranking system based on the likelihood of attacks to succeed, the value of targeted assets and the amount of interest in the type of attack.

All discussed approaches have the ability to priorities alerts, but they also have limitations, especially in the technical aspects of the methods adopted. They used multiple factors to estimate the risk, but do not consider different weightings based upon the importance of different decision factors.

3. The proposed Model

3.1. Overview

In [1] and [2] we have proposed the New System Alert Management for IDSs based on a stateful pattern matching algorithm, which can classify alerts by their importance and reduce number of false positives considerably. In order to improve our system, we propose in this paper more efficient method to prioritize alerts generated by IDSs by evaluating each risk. We assess the risk as a composition of indicators extracted from alerts itself and target assets, and then apply these results of the risk assessment to filter alerts produced by the IDS as **High Risk, Medium Risk** or **Low Risk.**

In our proposed system (Figure 1), we've used binary traffics files of KDD'99 which is used in our previous work. Snort [18] is used to produce alerts of KDD'99 dataset network traffics. Snort is an open source signature based IDS which gets KDD'99 online traffic and then generates alert log files; these files are entered into our proposed system as the inputs. A pattern matching algorithm is used to filter alerts and classify them to different form.

3.2. System architecture

Our approach encompasses three phases for processing events: **Pre-processing phase, Collection phase** and **Post- processing phase**. Each phase provides a level of abstraction to the following one. Figure 1 shows the three abovementioned units.

Pre-processing Phase: In this phase Snort analyzes KDD'99 binary traffic and generates alert files. These alert files are entered in our proposed model as Inputs.

Collection phase: We call this phase also "aggregation and normalization phase", For all data are received from Snort at one location. Aggregation aims at unify alerts in a single format on just one location. Normalization requires a parser familiar with the types and formats of generated alert from Snort after processing them. Snort list files contain information about all packets in KDD'99 dataset [19]. Using this method, we will be able to observe all alerts in the same format.

Post-processing phase: In this phase, once we have all the data in one place, we can implement mechanisms that will improve detection sensitivity and reliability.

In our model, we use three post-processing methods:

- **Classification and Filtering:** In this unit we extract the needed information, such as: Date, IP source, IP Destination, Attack name, etc., and we store them into a database. This information is extracted by parsing the alert file using Perl and regular expression as we will see in section 4.
- **Risk assessment:** Each alert is evaluated in relation to its attack type, and the target host. Several parameters make it possible to qualify the level of danger (Risk) of an alert. It is important to understand their significance in order to be able to manage correctly the alarms according to their level of importance (Table 1).
- **Prioritization:** We prioritize alerts received automatically after calculating the Risk. The priority of an alert depends on the architecture and inventory of the organization's devices. Prioritization is one of the most important steps in filtering alerts received from the Snort output. It means the evaluation of an alert's importance in relation to the organization's environment.

To calculate the Risk, we use the parameters described in Table 1. Each parameter has a value. For the Alert Priority and Alert Reliability, these values are stored in a MariaDB database. For device values, the security administrator must add all the organization's devices, including Servers, Firewalls, Switches, Access Points, Network Printers, etc., and must assign to each equipment a value between 1 and 5 according to the value and the criticality of the device. On the other hand, the other parameters are related to each type or classification of attack by Snort, these values are stored… in a MariaDB database for later use.

We calculate the risk using the three previous indicators by the following formula (3.7).

Alert Priority (P) = {1-5} **(3.1)**

Device Value (D) = {1-5} **(3.2)**

Alert Reliability (R) = {1-10} **(3.3)**

$$\textbf{Risk Assessment(RA)} = \frac{(\mathbf{P}) * (\mathbf{D}) * (\mathbf{R})}{\mathbf{X}} \quad \textbf{(3.4)}$$

The risk should be between 1 and 10 as we will see in Table 2, so the X value and obtained by calculating the Risk using the Maximum Value of each parameter, for example:

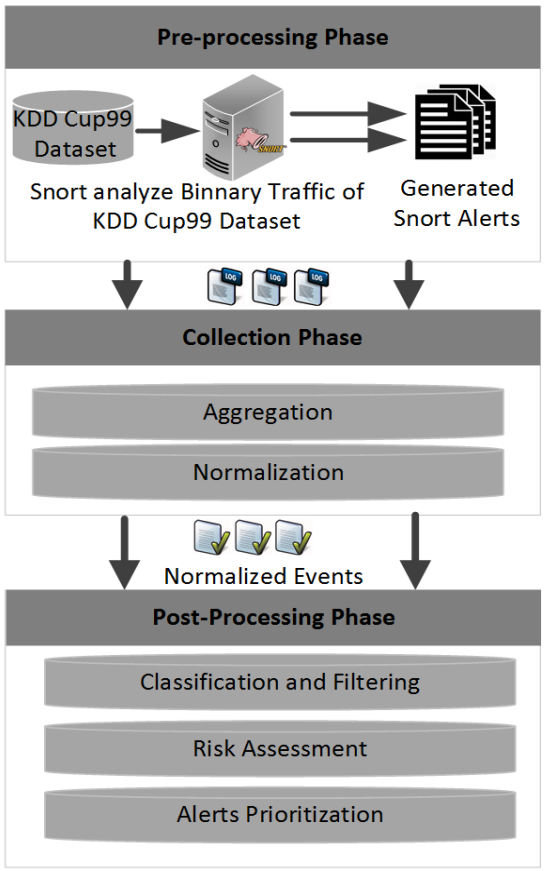


Figure 1: Proposed alert management system using real-time risk assessment and alert prioritization

Table 1. Description of parameters to assess the Risk of Alerts

Parameter	Description
Priority	Priority defines the order in which the action should be taken. A classification type rule assigns default priority that may be overridden with a priority rule [18]. In our model, we categorized priority into three types: **This should be Low = 1-2; Medium = 2-3; High = 4-5** These values are associated with each classification type of Snort IDS, and is stored into a MariaDB Database.
The value of the Destination Device associated with the event	This is a value to define the importance of a machine on the network. A DNS or Web server are more valuable resources for an organization than a network printer. As we will see later, these specifications will be taken into account when calculating the risk of each alert. This value must be between 1 and 5 (**1 = machine less important, 5 = very important machine**). This value is stored into a MariaDB Database for each device of the organization.
The likelihood that the event will occur **Reliability**	In terms of the risk, this parameter could be called "Reliability". This is defined for each independent event; an event may be a set of many alerts. The term reliability can be translated by the reliability that an event is not a *false positive*. The value of this parameter is between 1 and 10 **(equivalent to 1% =** this is a false **positive** and **100% =** it is not a false **positive).** This value is stored into a MariaDB Database and it is associated with an independent type of event (Alert Classification [18]).

Max(P)=5, Max(D)=5 and Max(R)=10

$$\mathbf{RiskAssessment(RA)} = \frac{\mathbf{5 * 5 * 10}}{\mathbf{X}} = \mathbf{10} \quad \textbf{(3.5)}$$

$$\mathbf{X} = \frac{\mathbf{250}}{\mathbf{10}} = \mathbf{25} \quad \textbf{(3.6)}$$

$$\mathbf{Risk\ Assessment(RA)} = \frac{\mathbf{(P) * (D) * (R)}}{\mathbf{25}} \quad \textbf{(3.7)}$$

The proposed model estimates the Risk for each alert. As we see in Table 1, the model uses a combination of three decision parameters.

Using the Risk Assessment, the Total Risk of an attack can be calculated according to the alert rate. Latter is calculated using formula (3.8) presented in [2], the Total Risk of an Attack (TRA) is used to evaluate the Risk of an attack type in a Meta-Alerts. Meta-Alerts can be generated for the clusters that contain all relevant information whereas the amount of data (i.e. alerts), for example all alerts to a specific host with the same attack type, or all alerts from the same source to the same destination and with the same attack type, etc., thus, the TRA is calculated using formula (3.9).

$$\mathbf{AR} = \frac{\mathbf{Number\ of\ Alerts\ byAttack}}{\mathbf{Tolal\ Alerts}} * \mathbf{100} \quad \textbf{(3.8)}$$

$$\mathbf{TRA} = \frac{\mathbf{(RA) * (AR)}}{\mathbf{100}} * \mathbf{10} \quad \textbf{(3.9)}$$

The resulting value can be mapped to the following Risk Categories, Table 2:

Table 2. Risk Assessment Categories

Risk Value	Signification
1-4	Low
5-7	Medium
8-10	High

3.3. Proposed Algorithm

The Algorithm can be explained as follows:

Algorithm: Filtering, classifying and prioritizing alerts according to the Risk Level

Input: Snort Log File (Generated by analyzing KDD'99 Dataset)
Output : Alerts with high Risk

1. Initializes the program
2. Processes the configuration and log files
3. While the number of alerts in log file is not reached
4. Extracts and records details of each alert into database
5. Correlate and classify alerts into many classes (attacks types)
6. Calculate the Alert Rate
7. For each alert in log file
8. Calculate the Risk of alert using (3.7)
9. End For
10. For each Meta-alert in log file
11. Calculate the Total Risk on an Attack TRA of Meta-Alerts using (3.9)
12. End For
13. Prioritize Alerts according to the Risk Assessment
14. Generate alarms if the Risk ≥ 70 %
15. End While

A basic flowchart diagram for the proposed algorithm is shown below (Figure 2):

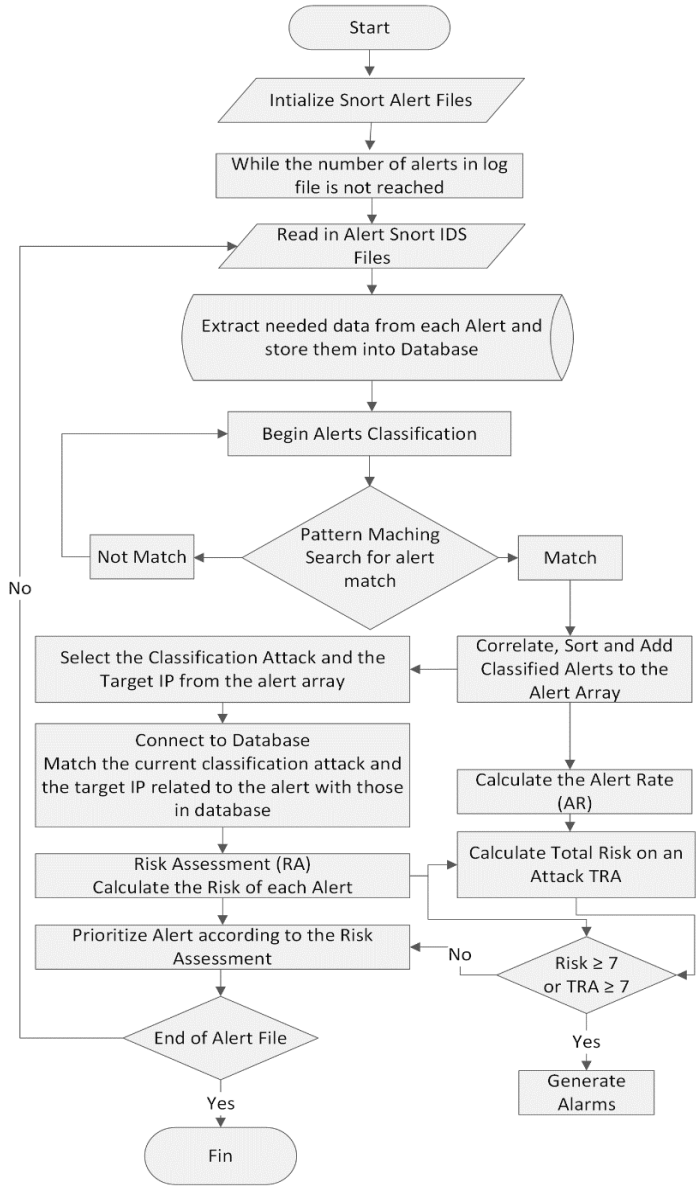


Figure 2: Flowchart of the proposed approach

4. Implementation and results

In order to test the effectiveness of our proposed model, we had it implemented and tested against the KDD'99 dataset. We also used Snort to scan the binary traffic of the dataset. Alerts generated by Snort were analyzed by our model using a pattern matching algorithm. To examine each alert, we wrote a Perl program using regular expression that parse the alerts and extract specific

parameters such as source IP, destination IP, attack name, priority etc... Finally, we use the formula (3.7) to assess the Risk of each alert and (3.9) to evaluate the Total Risk of attacks.

The example of alert below is just one sample of many alerts that we can find in one alert file. In this alert we can find many useful data, such as the Attack Name, Attack Classification, Target IP, Destination IP etc., to extract this useful information we use the Pattern matching rule below using Regular Expression (RE).

An example of alert to deal:

```
[**] [3:19187:7] PROTOCOL-DNS TMG Firewall Client long host entry exploit attempt [**] [Classification: Attempted User Privilege Gain] [Priority: 5] 03/19-16:01:43.762260 10.0.0.254:53 -> 172.16.2.11:1575 UDP TTL:64 TOS:0x0 ID:0 IpLen:20 DgmLen:201 DF Len: 173
```

These variables are stored into a MariaDB Database, thereby we can calculate the Risk for each generated Alert using these parameters and the formula (3.7), after that, using (3.9) we estimate the TRA.

Tables 3 and 4 below presents our experiment results using the output of Snort IDS and KDD'99 Dataset (with 99503 Alerts) that contains different 34 attempted attacks, 1056 Source IP, 485 Destination IP.

The Associated Pattern Matching rule:

```
^(?P<delimeter1>\[\*\*\])\s(?P<SigID_Rev>\[[0-
9:]+\])\s(?P<Attack_Name>[^\"]+)\s+(?P<delimeter2>\[\*\*\])\s\[Classification\:(?P<Attack_Class>[^\]]*)\]\s\[Priority\:(?P<Attack_Priority>[^\]]*)\]\s+(?P<Date>\d+\/\d+\-\d+\:\d+\:\d+)\.\d+\s(?P<SrcIP>\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})\:(?P<SrcPORT>\d+)\s\->\s(?P<DstIP>\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3})\:(?P<DstPORT>\d+)\s(?P<Protocol>\w+)\s+[^\]]*.*$
```

Extracting Data to variables:

```
Attack_Name = PROTOCOL-DNS TMG Firewall Client long host entry exploit attempt
Attack_Class = Attempted User Privilege Gain
Attack_Priority =
SrcIP = 10.0.0.254
DstIP = 172.16.2.11
DstPORT=53
Protocol= UDP
```

Tables 3. Classifying alerts by Detection Rate

Attack Name	Nr. of Events	Alert Rate %
(http_inspect) NO CONTENT-LENGTH OR+E3:E19 TRANSFER-ENCODING IN HTTP RESPONSE	45600	45.82 %
(http_inspect) INVALID CONTENT-LENGTH OR CHUNK SIZE	30524	30.67 %
(spp_sdf) SDF Combination Alert	12356	12.41 %
Consecutive TCP small segments exceeding threshold	7052	7.08 %
(http_inspect) UNESCAPED SPACE IN HTTP URI	1159	1.16 %
(http_inspect) LONG HEADER	809	0.81 %
(http_inspect) SIMPLE REQUEST	640	0.64 %
ET CHAT IRC PRIVMSG command	425	0.42 %
ET CHAT IRC PING command	276	0.28 %
ET CHAT IRC PONG response	127	0.12 %
ET CHAT IRC USER command	101	0.10 %
ET CHAT IRC NICK command	99	0.10 %
ET CHAT IRC JOIN command	84	0.08 %
ET POLICY Outbound Multiple Non-SMTP Server Emails	74	0.07 %
Reset outside window	35	0.03 %
(http_inspect) UNKNOWN METHOD	29	0.03 %
(ftp_telnet) FTP bounce attempt	16	0.01 %
ET SCAN Potential SSH Scan OUTBOUND	15	0.01 %
(http_inspect) NO CONTENT-LENGTH OR+E3:E19 TRANSFER-ENCODING IN HTTP RESPONSE	12	0.01 %
(http_inspect) INVALID CONTENT-LENGTH OR CHUNK SIZE	11	0.01 %
(spp_sdf) SDF Combination Alert	11	0.01 %
Consecutive TCP small segments exceeding threshold	11	0.01 %
(http_inspect) UNESCAPED SPACE IN HTTP URI	8	0.01 %
(http_inspect) LONG HEADER	6	0.00 %
(http_inspect) SIMPLE REQUEST	5	0.00 %
ET CHAT IRC PRIVMSG command	5	0.00 %
ET CHAT IRC PING command	3	0.00 %
ET CHAT IRC PONG response	3	0.00 %
ET CHAT IRC USER command	3	0.00 %
ET CHAT IRC NICK command	2	0.00 %
ET CHAT IRC JOIN command	2	0.00 %

Our platform for the experiment are described as follows:

- Processor: Intel (R) Core (TM) i7-6500U CPU @ 2.50GHZ 2.59 GHZ.
- Memory: 4 GB
- System (OS): Linux Ubuntu Server 16.04 64-bit

Table 3 present the classification of alerts according to the alert Rate; Alert Rate is measured by calculating the number of alert for each attack [2].

Table 4 shows the process of evaluating the risk of alerts. After classifying them into three categories, the security administrator will see just alerts with medium and high risk. The rest with low risk can be considered as false positives, this is related to the values of the target hosts and the reliability and priority of the attack. As we can see in Table 4, the risk is evaluated for each alert. For example, in the first row in the table, we notice that the attack "**(http_inspect) NO CONTENT-LENGTH OR+E3:E19 TRANSFER-ENCODING IN HTTP RESPONSE**", has been generated **45600** times to the target "**192.168.11.52**". The **RA** is **10** which is **High**. It is evaluated using the three parameters related to the alert: the priority, the reliability and the device target value. In other ways, the Total Risk of an Attack is evaluated according to the number of events of each attack in relation to the Alert Rate of this Attack.

The specific and complex characteristics of the network system environment make the implementation of Intrusion Detection System more difficult with the multitude of alerts and the huge number of false positives. Therefore, the new approach for detection and analysis of malicious activities is needed in order to check the effectiveness of the current security controls that protect information data.

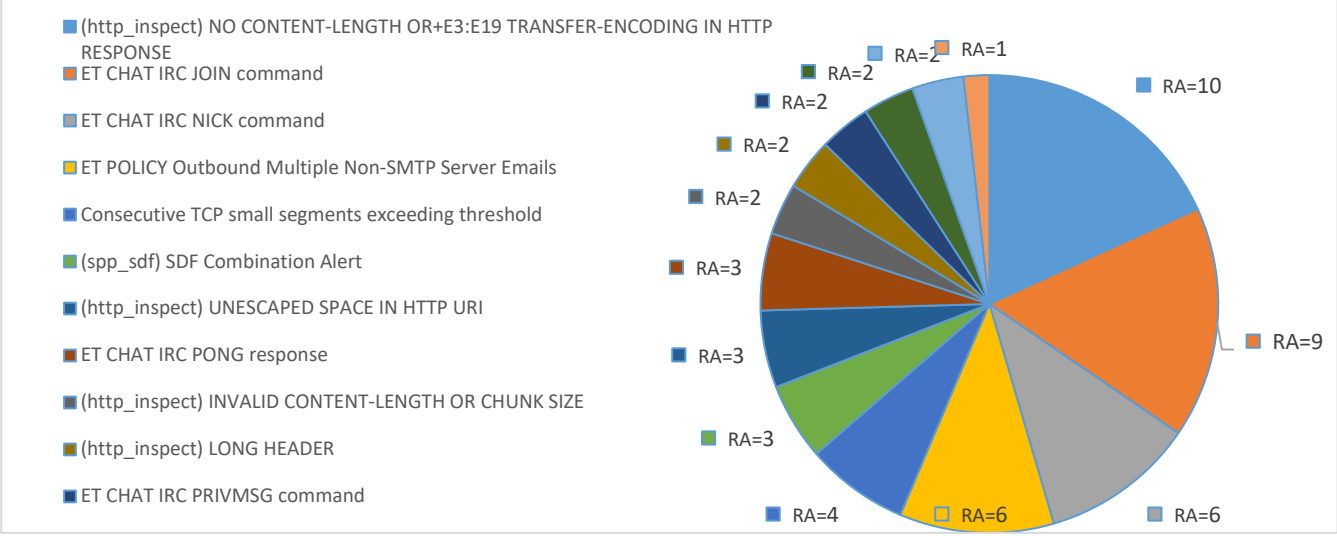


Figure 3: Alerts prioritization after risk assessment

Table 4. Risk Assessment according to the alert and target host parameters

No. of events	To Target	Attack Name	P	R	D	RA	RA Ranking	AR %	TRA %
45600	192.168.11.52	(http_inspect) NO CONTENT-LENGTH OR+E3:E19 TRANSFER-ENCODING IN HTTP RESPONSE	5	10	5	10	**High**	45.82 %	**4.6%**
30524	192.168.2.12	(http_inspect) INVALID CONTENT-LENGTH OR CHUNK SIZE	3	5	4	2	**Low**	30.67 %	**0.6%**
12356	192.168.11.5	(spp_sdf) SDF Combination Alert	3	5	5	3	**Low**	12.41 %	**0.4%**
7052	172.16.2.56	Consecutive TCP small segments exceeding threshold	3	7	5	4	**Low**	7.08 %	**0.3%**
1159	192.168.2.100	(http_inspect) UNESCAPED SPACE IN HTTP URI	3	7	3	3	**Low**	1.16 %	**0.03%**
809	172.16.16.123	(http_inspect) LONG HEADER	3	5	3	2	**Low**	0.81 %	**0.01%**
640	172.16.16.52	(http_inspect) SIMPLE REQUEST	3	5	2	1	**Low**	0.64 %	**0.06%**
425	192.168.11.101	ET CHAT IRC PRIVMSG command	3	5	4	2	**Low**	0.42 %	**0.008%**
276	192.168.2.58	ET CHAT IRC PING command	3	5	4	2	**Low**	0.28 %	**0.005%**
127	10.10.101.2	ET CHAT IRC PONG response	3	7	4	3	**Low**	0.12 %	**0.003%**
101	10.222.1.1	ET CHAT IRC USER command	3	5	4	2	**Low**	0.10 %	**0.002%**
99	10.22.1.231	ET CHAT IRC NICK command	5	8	5	6	**Medium**	0.10 %	**0.006%**
84	172.16.2.112	ET CHAT IRC JOIN command	5	9	5	9	**High**	0.08 %	**0.007%**
74	172.16.16.52	ET POLICY Outbound Multiple Non-SMTP Server Emails	5	8	4	6	**Medium**	0.07 %	**0.004%**

In this work we demonstrated how the implementation of risk assessment reduces the number of false positive. With such an approach, the security network administrator will see just the alerts with a medium and high-risk level that presents a real threat to the organization. The rest of alerts will be considered as a false positive and will not be sent as alarms. Thus, the network security administrator can check the effectiveness of the current security controls that protect the organization's assets (Figure 3). In the next step of this work, we will set up a knowledge base for all the false positives to be compared further and to see if they are indeed.

5. Conclusion and future works

A novel approach that evaluates intrusion detection system alerts using a new risk assessment and alert prioritization is presented. We proposed a model that recognizes and analyzes malicious actions by calculating the risk related to attack pattern and qualify the level of dangerousness of each attack, thus prioritizing alerts generated by IDS. The implementation has demonstrated the efficiency of our approach in both decreasing the huge number of false positives that can reach over 95 % of alerts in the usual cases with a normal IDS, using our model we can control the rate of false positives; thus, we increase the effectiveness of the IDS system.

In the next step of this work, we will focus on the implementation of our approach with other IDSs. Moreover, we will improve our model by using advanced functions as well as more sophisticated algorithms such as machine learning algorithms to classify the attacks according to their dangerousness.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] E.M. Chakir, M.Moughit, Y.I. Khamlichi, "An efficient method for evaluating alerts of Intrusion Detection Systems" in 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 2017. Pages: 1 - 6, https://doi.org/10.1109/WITS.2017.7934678

[2] E.M. Chakir, Y.I. Khamlichi, M.Moughit, "Handling alerts for intrusion detection system using stateful pattern matching" in 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), 2016, Pages: 139 - 144, https://doi.org/10.1109/CIST.2016.7805031

[3] M. Ahmed, A.N. Mahmoud,J. Hu, "A survey of network anomaly detection techniques", 2016. J. Netw. Comput. Appl. 60, 19– 31.

[4] N. Moustafa, J. Slay, " The evaluation of network anomaly detection systems: statistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set", Perspect. 25 (1 –3), 18– 31. 2016. Inf. Secur. J.: a Glob.

[5] A. Shameli-Sendi, M. Chariet, A. Hamou-lhadj, "Taxonomy of Intrusion Risk Assessment and Response System", ELSEVIER Volume 45, , Pages 1–16. September 2014.

[6] H.M. Bhuyan, D.K. Bhattacharyya, JK. Kalita, "An effective unsupervised network anomaly detection method" In International conference on advances in computing, communications and informatics, no. 1; p. 533–9. 2012.

[7] S. Wang, Z. Zhang, and Y. Kadobayashi," Exploring attack graph for cost-benefit security hardening: A probabilistic approach," Computers & Security, vol. 32, pp. 158-169, 2013.

[8] Zhang, Z., Ho, P. and He, L. "Measuring ID S-estimated attack impacts for rational incident response: A decision theoretic approach", Computers & Security , Vol. 28 No. 7, pp. 605-614, 2009.

[9] K. Scarfone and P. Mell: Guide to Intrusion Detection and Prevention Systems. Technical report, NIST: National Institute of Standards and Technology, U.S. Department of Com-merce, 2007.

[10] Hung-Jen Liao, Ch.-H. Richard Lin, Y.-Ch. Lin, K.-Y. Tung,"Intrusion detection system: A comprehensive review", Journal of Network and Computer Applications 36, pp: 16–24. 2013.

[11] N. B. Anuar, H. Sallehudin, A. Gani, and O. Zakaria," Identifying False Alarm for Network Intrusion Detection System Using Hybrid Data Mining and Decision Tree," Malaysian Journal of Computer Science, ISSN 0127- 9084, pp. 110-115. 2008.

[12] H. Debar, A. Wespi, "Aggregation and correlation of intrusion-detection alerts" , Recent Advances in Intrusion Detection, 2001.

[13] N. B. Anuar, H. Sallehudin, A. Gani, and O. Zakaria," A risk index model for security incident prioritisation" Proceedings of the 9th Australian Information Security Management Conference, Edith Cowan University, Perth Western Australia, 5th -7th December, 2011.

[14] Alsubhi, K., Al-Shaer, E. and Boutaba, R. "Alert prioritization in intrusion detection systems", Proceedings of the IEEE Network Operations and Management Symposium, Salvador, Brazil, pp. 33-40, 2008.

[15] Dondo, M.G., "A vulnerability prioritization system using a fuzzy risk analysis approach", Proceedings of the 23rd International Informa tion Security Conference, Milano, Italy, pp. 525-539, 2008.

[16] Lee, B. Chung, H. Kim, Y. Lee, C. Park, and H. Yoon," Real-time analysis of intrusion detection alerts via correlation," Computers & Security, vol. 25, no. 3, pp. 169-183, 2006.

[17] P.A. Porras, M.W. Fong, and A. Valdes, "A Mission-Impact-Based Approach to INFOSEC Alarm Correlation, Proceedings of the 5th International Symposium on Recent Advances in Intrusion Detection (RAID 2002)" 95–114, 2002.

[18] The Snort Project, Snort user's manual https://www.snort.org/downloads, 2017.

[19] KDDCup '99 dataset, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 25-29 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Virtual Memory Introspection Framework for Cyber Threat Detection in Virtual Environment

Himanshu Upadhyay*, Hardik Gohel, Alexander Pons, Leo Lagos

Applied Research Center, Florida International University, Miami, 33172, United States

ARTICLE INFO

Article history:
Received: 01 November, 2017
Accepted: 13 December, 2017
Online: 18 January, 2018

Keywords :
Cybersecurity,
Virtualization,
Linux threat detection,
Hypervisor,
Feature selection technique,
Memory forensic analysis,
Virtual machine introspection

ABSTRACT

In today's information based world, it is increasingly important to safeguard the data owned by any organization, be it intellectual property or personal information. With ever increasing sophistication of malware, it is imperative to come up with an automated and advanced methods of attack vector recognition and isolation. Existing methods are not dynamic enough to adapt to the behavioral complexity of new malware. Widely used operating systems, especially Linux, have a popular perception of being more secure than other operating systems (e.g. Windows), but this is not necessarily true. The open source nature of the Linux operating system is a double edge sword; malicious actors having full access to the kernel code does not reassure the IT world of Linux's vulnerabilities. Recent widely reported hacking attacks on reputable organizations have mostly been on Linux servers. Most new malwares are able to neutralize existing defenses on the Linux operating system. A radical solution for malware detection is needed – one which cannot be detected and damaged by malicious code. In this paper, we propose a novel framework design that uses virtualization to isolate and monitor Linux environments. The framework uses the well-known Xen hypervisor to host server environments and uses a Virtual Memory Introspection framework to capture process behavior. The behavioral data is analyzed using sophisticated machine learning algorithms to flag potential cyber threats. The framework can be enhanced to have self-healing properties: any compromised hosts are immediately replaced by their uncompromised versions, limiting the exposure to the wider enterprise network.

1. Introduction

Dependency on computer systems has been growing exponentially in recent years. Government offices and business organizations are major targets of malicious actors for stealing highly valuable data. These entities are major Linux adopters nowadays because of its reputed safety and security. They are attempting to protect themselves from cyberattacks with digital defense techniques like encryption, firewalls and heuristic or signature scanning packages. Meanwhile, the number of attacks that involve infiltrating military data centers, targeting power grids, and stealing trade secrets from both private and public organizations continues to increase. The detection, response and reporting of these kinds of intrusions as well as other incidents involving computer systems, are crucial for cybersecurity professionals. This paper proposes a Virtual Memory Introspection based framework for detection, analysis and monitoring of malware behavior using memory forensics, during cyberspace attacks in a virtualized environment. The framework provides advanced instrumentation tools for control and monitoring of malware, fine-grained introspection of operating system Kernel and user process behavior using the well-known LibVMI technology, and analysis of behavioral data using state-of-art machine learning techniques. In study of research literature, one of the major methods of malware detection that has emerged over the years is Linux memory forensics [1] [2]. A number of authors have described novel detection systems using Memory Forensics, or using kernel data structure invariants as a reference frame to identify rootkit intrusions [3] [4]. The goal of this test technology is to facilitate threat assessment of malware, to understand its goals, and degrade impact on the compromised systems [5].

Further, as cyber-attacks continues to expand and the sophistication of the adversaries grows, defenders must adapt quickly in order to survive. There are numerous kinds of malware prevalent today, including Backdoors, Bots, Downloaders and Droppers, Ransomware, Rootkits, Scareware, Spyware, Exploits, Logic Bombs, Trojans, Viruses, Worms, etc. The detection,

*Dr. Himanshu Upadhyay, Email: upadhyay@fiu.edu

response and reporting of these kinds of intrusions as well as other incidents involving computer systems, are crucial for cybersecurity professionals. Present static methods used by anti-viruses are insufficient to stop advanced malware threats, which are capable of disabling the anti-viral software via root-kit mechanisms.

2. Virtualization & Data Analytics Framework

The proposed framework has the following major components as shown in Figure1:

2.1. Virtualization

The virtual machines, which are monitored for malware attacks, are collectively called the system under test (SUT). This system is hosted on a Xen hypervisor based virtualization platform. The behavioral data on the guest VMs is captured through Virtual Memory Introspection (VMI) using the LibVMI framework

2.2. Data Analytics

This platform consists of various traditional machine learning algorithms used to train the models and perform prediction using various test vectors consisting of malware and rootkits on the virtual machines (SUT). In this proposed framework, the authors are building an advanced data analytics platform on the database server, R runtime with in-memory analytics providing the scale and performance. Various traditional machines learning algorithms like Random Forest, Support Vector Machines, Logistics regression etc. will be utilized to perform Data Analytics. This platform analyzes the memory data structure captured on the VMI framework using cutting-edge machine learning algorithms. These algorithms are used to build the model to predict the malware behavior and display the results.

2.3. Test Control Center

The Test Control Center is an application that provides control and administration of the entire framework with an intuitive web based interface. The operator can create virtual machines (SUT), install benign applications and malware, capture behavioral information and use the Data Analytics platform to build the model and test virtual machines for malware using various traditional machine learning algorithms. Test Control Center is a centralized application to manage the test bed and execute various test cases. In the proposed framework, the authors are developing various modules like virtual machine management, network map, test case management, model management, configuration, testbed administration and help.

The three platforms working together will enable an organization to detect malware almost in real time and monitor its behavior in a virtual environment. The user provides a set of well-defined inputs to the virtualization platform based on predetermined test cases, and captures the kernel data structures information and transfers it to the Data Analytics platform. The Data Analytics platform uses traditional machine learning algorithms to build the model and predicts results based on the information extracted from guest virtual machine memory and displays the results on the Test Control Center. The following diagram in Figure 1 provides the overall system design of the framework:

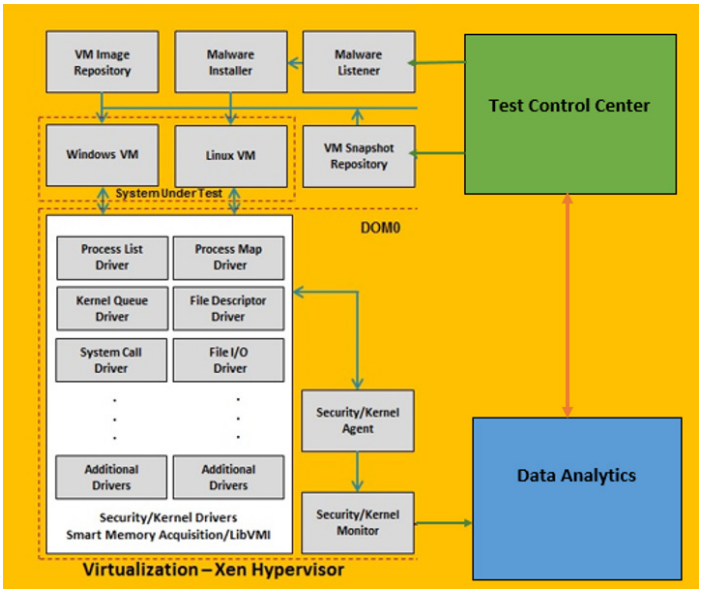


Figure 1: Virtualization & Data Analytics Framework System Diagram

The various components of the virtualization platform are discussed below.

I. VM Image Repository

Images of the different virtual machines will be stored in this location. It provides the user the ability to select the virtual machine images that will create the virtual machines for SUT with different configurations.

II. Malware Listener and Installer

This module listens to the command from the malware management module and communicates with the malware installer on the virtual machine to install the selected malware.

III. VM Snapshot Repository

This module will store the snapshots of normal operation which will later be used to roll back to the clean state after the test execution.

IV. Multiple Linux VMs

These are the guest virtual machines created by the operator based on predefined configurations.

V. Kernel Agent

Security/kernel agent manages all the drivers to extract the kernel data structures from the SUT. They are deployed on DOM 0 which is the secured location on the hypervisor with required privileges.

VI. Kernel Drivers

These the individual drivers which are used for smart memory acquisition of the kernel data structures from the SUT. These drivers use the LibVMI libraries for virtual memory introspection from the Dom0 level to extract kernel data structures. They are managed by Security/Kernel Agent.

VII. Kernel Monitor

This module communicates with data analytics platform and sends the data extracted by the Security/kernel agent to the listener on the data analytics platform to store the data in to the database.

VIII. Kernel Agent Listener

The Agent Listener plays a crucial role in the virtualization framework. This module communicates with the security/kernel monitor in the virtualization framework, and receives the extracted data from the security/kernel agent and stores it into the central database.

3. Virtual Memory Introspection (VMI) Framework

The VMI platform of this framework is built on the "Xen Hypervisor" which enables the operator to create multiple virtual machines with different configuration and perform system testing with the environment. The Linux machines are used as SUT, to be monitored for malware intrusion are hosted on this hypervisor. Xen hypervisor [6] provides two security domains for the hosted virtual machines. The virtual machine running under Dom0 (Domain 0) on the Hypervisor controls the resource allocation for the virtual machines that runs the Linux Guest operating system [7].

The major key objectives of virtualization framework are listed below:

I. Develop innovative ways to detect and monitor malware in a virtualized testbed with smart memory acquisition.

II. Deployment of virtualized environments along with advanced tools developed through this research for control and monitoring of the cyberspace test environment.

III. Cyber-attack emulation through infection and propagation of simulated endpoints.

IV. Virtual machine introspection, data collection, and monitoring of various aspects of the infrastructure through a centralized system.

V. Display results on the Test Control Center to monitor the impact of malware on SUT.

The SUT can be Linux or Windows guest virtual machines. Using the Test Control Center, the operator is able to create guest VMs on the Xen hypervisor from a virtual machine repository, introduce pre-determined malware into them, and capture kernel data structure information and stores them into the central database. We have developed an application called "Kernel Agent" that runs on the Dom0 virtual machine, and uses kernel drivers to perform smart memory acquisition on the guest operating system. This kernel information, which includes process details, memory data structures, and file system information, is written to a central database for data analytics.

Xen hypervisor is used to host, configure, and control the guest Linux virtual machine that will be tested for malware. The introspection is carried out through the "LibVMI" framework in conjunction with Google's "Rekall" profiles. The Rekall profile is a JSON file that contains the address mappings of all Linux Kernel data structures. By integrating LibVMI with Google Rekall, the process of extracting Kernel configuration parameters and data structures at run-time can be automated. This facilitates the smart memory acquisition of process behavioral data from the kernel. We are using Google's "Go" programming language for writing the introspector application, and to push the acquired data to the central database. The Go programming language can integrate with suitable C/C++ libraries, and significantly reduces the time needed for development and testing new build of each module. It provides better control over distribution and parallelization mechanisms of the Security Agent. We are using "Libvirt" library for management of the guest virtual machines through a custom built Introspector. Introspection requests are serviced by the LibVMI library, while the Libvirt library is used to facilitate the creation, starting, stopping, pausing and resuming of the guest Linux virtual machines [8].

The Introspector is comprised of several core subsystems that are necessary for lifecycle management and introspection of virtual machines on a host. The Introspector also maintains two socket connections that listen to requests for introspection or virtual machine administration coming from the Test Control Center.

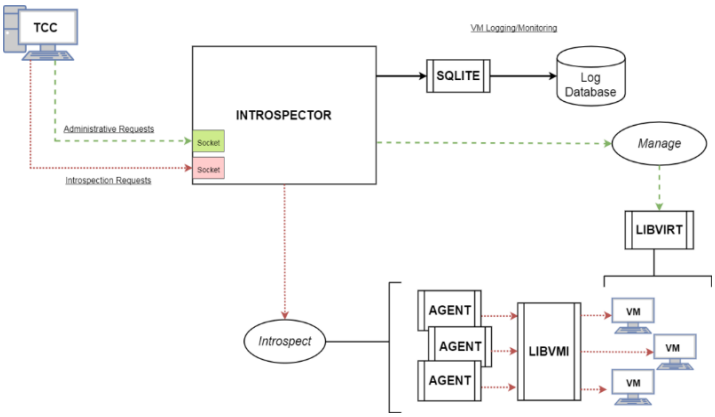


Figure 2: Proposed Virtual Machine Introspection Framework

Introspector: startup

Introspector (introspector.go) builds a Settings object (settings.go) by reading the values in the introspector.conf settings file. The configuration file is used to determine which IP/Port combination to use for handling virtual machine administration requests and introspection requests. The Introspector will also start the StateManager that manages a SQLite database containing the current state of the virtual machines in the system.

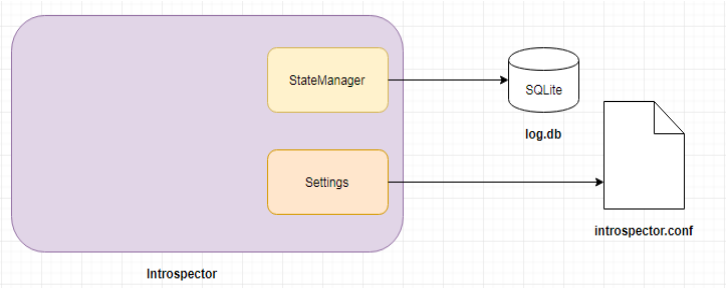


Figure 3: Introspector setup with Database

Next, a thread is spawned that opens a socket connection to listen for incoming requests for virtual machine administration from the Test Control Center while the main Introspector thread waits on another socket for incoming introspection requests. Administration requests are XML requests that have the root node <LibvirtServerMessage>

4. Research on Kernel Data Structures

The analysis conducted on the footprint of a process in kernel space is in near real-time, to distinguish between benign and malicious processes. The Linux Kernel maintains data structures that contain information about every action and resources used by a process. The study thus far has identified 15 features from 118 features which can effectively distinguish between benign and malicious processes [9] [10]. The next steps in the pipeline is to perform virtualization to extract threads, system call , invariant data structures like system call table and interrupt descriptor table, IP addresses, network sockets, URLs, open files, passwords, catches, clipboards and other user generated content, encrypted keys, and configurations of hardware and software [11].

5. Benign and Malware Applications

We have also performed research to extract the kernel data structures from the Linux kernel task structure using LibVMI library. The following Benign processes and Malwares samples have been identified to be installed on the Linux virtual machines. The Linux kernel version under consideration as of the writing of this paper is 3.16.0-23- generic.ko in Ubuntu 14.04 [11] [12].

GUI Applications	CLI Applications
1) Firefox - Web Browser 2) Thunderbird - Email client application 3) gpaint - Paint Application 4) Libre - Office writer 5) Rythmbox - Music player 6) Connectagram - Word unscrambling game 7) Arora - Cross platform web browser 8) Empathy- Internet messaging application 9) Alarm clock 10) Shotwell- Photo manager	1) Alpine-Pico- Is a text editor, uses the pine email client for writing email messages. Run from the terminal as pico, pico.alpine 2) Aaphoto- An image manipulation tool for automatic color correction of photos 3) ACL2- Programming language in which user can model computer systems and a tool to help prove properties of those models. 4) Python- Programming language 5) Gedit Text editing tool 6) Calcoo - scientific calculator 7) Calcurse- text based calendar and todo manager 8) Clamav-daemon- antivirus utility for unix-scanner daemon run from the terminal 9) Gzip- Zipping application

6. Linux Kernel Data Structure Extraction

At the outset we are going to extract the following task structure list from Linux data structures. Based on our research, the following list of features has been identified for extraction to test the virtualization framework in the pilot stage. These are the primary features of the processes that will run in the kernel of the Linux virtual machine at runtime [13] [14].

Sr. No.	Features' Name	Description
1	map_count	Number of memory regions of a process
2	page table lock	Used to manage the page table entries
3	hiwater rss	Number of page frames that a process owns
4	shared_vm	Number of pages in shared file memory mapping of a process
5	exec_vm	Number of pages in exec. memory mapping of process
6	nr_ptes	Number of page tables owned by a process
7	utime	Execution time of a process in user mode (tick count)
8	stime	Execution time of a process in kernel mode (tick count)
9	nvcsw	Volunteer context switches of a process
10	nivcsw	Involuntary context switches
11	total_vm	Size of process's address space in terms of Number of pages
12	min_flt	Minor page faults of a process
13	alloc lock.raw lock.slock	Used to lock memory manager, files and file system etc.
14	hiwater_vm	Max Number of pages appeared in memory region of process
15	fs.count	fs_struct's usage count to indicate the restrictions

7. Data Analytics and Results

The research on data analytics is currently on-going [15] [16]. The Data Analytics platform consists of centralized database servers for analytics and processing. The research will be focused on identifying the classification and anomaly detection machine learning algorithms that includes open source and commercial platforms, libraries with these platforms [17] [18] to detect malware [19] [20] that primarily focus on the Linux data structure extracted by the VMI Framework.

Once the framework is narrowed down and tuned as a viable solution for malware detection, the results will be summarized to demonstrate which machine learning models are powerful for malware detection and monitoring. The research outcome will consist of three components: memory data structures, algorithms with pros/cons and machine learning model performance data.

8. Conclusion and Future Work

The goal of the VMI framework research is to provide solid foundation for security of the Linux operating system as well as the capabilities to identify and monitor cyber threats in virtual environment. The key challenges of this research is to identify various data structures affected by modern malware, and in-depth

examination of various machine learning algorithms with memory forensics to solve key cybersecurity issues.

The focus of this research is to develop a system that can work with the latest Linux operating systems providing memory forensics capability in a virtual environment. In the future, we plan to explore all possible ways of data analytics with big data technologies and deep learning. With sufficient time and effort, the development of this framework can result in an extremely powerful system for early threat detection and defense.

Acknowledgment

This research is supported by Department of Defense (DOD) – Test Resource Management Center (TRMC), USA. We thank our colleagues who provided insight and expertise, directly or indirectly, that greatly assisted the research.

References

[1] M H Ligh, A Case, J Levy, A Walters. "The Art of Memory Forensics", 2014
[2] M Wade, "Memory Forensics: Where to Start" at http://www.forensicmag.com/ article/2011/06/ memory-forensics-where-start, 2011
[3] Baliga, A., Ganapathy, V. and Iftode, L., Detecting kernel-level rootkits using data structure invariants. IEEE Transactions on Dependable and Secure Computing, 8(5), pp.670-684, 2011.
[4] D. Levy, H. A. Gohel, H. Upadhyay, A. Pons and L. E. Lagos, "Design of Virtualization Framework to Detect Cyber Threats in Linux Environment," 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), New York, NY, pp. 316-320, 2017
[5] H. Gohel. "Introduction to Network & Cyber Security", 2015
[6] Xen Project available at https://www.xenproject.org/, 2013
[7] H. Gohel. "Looking Back at the Evolution of the Internet." CSI Communications - Knowledge Digest for IT Community Vol. 38 Issue. 6 pp. 23-26, 2014
[8] N Joshi, D.B.Choksi, "Implementation of process forensic for system calls", International Journal of Advanced Research in Engineering and Technology (IJARET), ISSN 0976 – 6480(Print), ISSN 0976 – 6499(Online) Vol. 5, Issue. 6, pp. 77-82, 2014
[9] Blackbag Team, "MEMORY FORENSICS", at https://www.blackbagtech.com/blog/2016/03/07/ windows-memory-forensics/2016
[10] Farrukh Shahzad, M. Shahzad, Muddassar Farooq, "In-execution dynamic malware analysis and detection by mining information in process control blocks of Linux OS" Information Sciences, Volume 231, pp. 45-63, ISSN 0020-0255,2013
[11] N Joshi, D. B. Choksi, "Implementation of Process Forensic for System Calls", International Journal of Advanced Research in Engineering & Technology (IJARET), Vol. 5, Issue. 6, pp. 77 - 82, ISSN Print: 0976-6480, ISSN Online: 0976-6499, 2014.
[12] H. Upadhyay, H. Gohel "Security Corner: Cyber Threat Analysis with Memory Forensics", CSIC – Knowledge Digest for IT Community, Vol. 40, Issue. 11, ISSN 0970-647X, pp. 17-19, 2017
[13] H. Upadhyay, H. Gohel "Design of Advanced Cyber Threat Analysis Framework for Memory Forensics", International Journal of Innovative Research in Computer and Communication Engineering, ISSN (Online): 2320-9801, ISSN (Print): 2320-9798, Vol. 5, Special Issue 2. pp.132-137, 2017
[14] F. Shahzad, S. Bhatti, M. Shahzad and M. Farooq, "In-Execution Malware Detection Using Task Structures of Linux Processes," 2011 IEEE International Conference on Communications (ICC), Kyoto, pp. 1-6, 2011.
[15] Z. Gu, Z. Deng, D. Xu and X. Jiang, "Process Implanting: A New Active Introspection Framework for Virtualization," IEEE 30th International Symposium on Reliable Distributed Systems, Madrid, pp. 147-156, 2011
[16] Hal Pomeranz, "Detecting Malware with Memory Forensics", at http://www.deer-run.com/~hal/ Detect_Malware_w_Memory_Forensics.pdf, 2015
[17] Hizver, Jennia, and Tzi-cker Chiueh. "Real-time deep virtual machine introspection and its applications." ACM SIGPLAN Notices. Vol. 49. No. 7. ACM, 2014.
[18] Hardik, Gohel. "Data Science - Data, Tools & Technologies." CSI Communications Knowledge Digest for IT Community, Vol. 39 Issue. 3, pp. 8-10, 2015
[19] H Gohel, P Sharma. "Study of Quantum Computing with Significance of Machine Learning." CSI Communications - Knowledge Digest for IT Community, Vol. 38, Issue. 11, pp. 21-23, 2015
[20] E Mariconti, O Lucky, P. Andriotis, "Detecting Android Malware by Building Markov Chains of Behavioural Models", NDDS'17, San Diego, USA, 2017

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 30-37 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

A Test Code Generation Method for Coding Standard Input/Output with Exception Handling in Java Programming Learning Assistant System

Ei Ei Mon[1], Nobuo Funabiki[*,1], Ryota Kusaka[1], Khin Khin Zaw[1], Wen-Chung Kao[2]

[1]*Okayama University, Department of Electrical and Communication Engineering, Okayama, Japan*

[2]*National Taiwan Normal University, Department of Electrical Engineering, Taipei, Taiwan*

ARTICLEINFO

Article history:
Received: 30 October, 2017
Accepted: 13 December, 2017
Online: 30 January, 2018

Keywords:
Java programming
JPLAS
Test code
Test case
Automatic generation
JUnit

ABSTRACT

To advance Java programming educations, we have developed the Java Programming Learning Assistant System (JPLAS) that provides the code writing problem. This problem asks a student to write a source code to satisfy the specification of a given assignment, where the correctness is verified by running test code on JUnit. For a novice student, a code of implementing the standard input/output with the exception handling should be mastered at the early stage as the first step programming for human interfaces. However, for a teacher, it is not easy to write the test code for it. In this paper, we propose a test code generation method to generate the test code using the reference source code for the assignment. In the evaluation of this proposal, all the students completed the codes using the generated test codes for exception handling, although the use of exception handling functions was sometimes insufficient or incorrect.

1 Introduction

Recently, the objected oriented programming language *Java* has been widely used in various practical application systems in societies and industries due to the high reliability, portability, and scalability. Java was selected as the most popular programming language in 2015 [2]. Therefore, there have been strong demands from industries for Java programming educations. Correspondingly, a plenty of universities and professional schools are currently offering Java programming courses to meet this challenge. A typical Java programming course consists of grammar instructions in the class and programming exercises in computer operations.

To advance Java programming educations, we have developed the Web-based *Java Programming Learning Assistant System (JPLAS)* [3]-[7]. JPLAS inspires students by offering sophisticated learning environments via quick responses to their answers for self-studies. At the same time, it supports teachers by reducing loads of evaluating codes. JPLAS has several types of problems to cover a variety of students at different learning levels. Among them, the *code writing problem* [4] asks a student to write a source code to satisfy the specification of a given assignment.

The *code writing problem* is implemented based on the *test-driven development (TDD) method* [8], using an open source framework *JUnit* [9]. *JUnit* automatically tests the codes on the server to verify their correctness using the *test code* when they are submitted by students. Thus, students can repeat the cycle of writing, testing, modifying, and resubmitting codes by themselves, until they can complete the correct codes for the assignments.

To register a new assignment for the code writing problem in JPLAS, a teacher has to prepare a *problem statement* describing the code specification, a *reference source code*, and a *test code* using a Web browser. It is noted that the reference source code is essential to verify the correctness of the problem statement and the test code. Then, a student should write a source code for the assignment while referring the statement and the test code, so that the source code can be tested by using the given test code on *JUnit*.

However, teachers at schools are not accustomed to writing a test code that can run on *JUnit*. A teacher may spend much time in struggling to write a test code, and may

[*]Nobuo Funabiki, Department of Electrical and Communication Engineering, Okayama University, Okayama, Japan, funabiki@okayama-u.ac.jp
This paper is an extension of work originally presented in 31st IEEE International Conference on Advanced Information Networking and Applications (AINA-2017) [1]

https://dx.doi.org/10.25046/aj030105

register an incomplete test code that does not verify some requirements described in the problem statement correctly. This incomplete test code must be avoided because it may produce inappropriate feedback to a student and undermine confidence to JPLAS. On the other hand, a commercial tool for generating a test code is usually expensive, and may not cover a test code that verifies the standard input/output with exception handling in a source code. The code of implementing the standard input/output with exception handling should be mastered by novice students at the early stage of Java programming educations as the first step programming for human interfaces.

In this paper, we propose a *test code generation method* for the code writing problem in JPLAS that generates a test code using a reference source code to test the standard input/output with exception handling. This method can generate a test code through the following steps: 1) a *test code template* is provided by our proposal, 2) a set of standard inputs to be tested are made by a teacher, 3) by running the reference code with each input, the corresponding expected standard output is extracted correctly, and 4) this pair of the standard input and the standard output are embedded into the test code template. By repeating steps 3) and 4) for every test standard input, the test code can be completed. To run the source code using the test code on *JUnit*, it introduces the classes to handle the standard input/output functions as the memory access functions in [10].

To evaluate the proposed method, first, we applied it to 97 source codes in Java programming textbooks or Web sites that contain the standard input/output. It has been proved that the generated test codes could correctly verify the source codes except for one code using a random generator. Then, we generated the test codes for three problems and asked five students who are currently studying Java programming to write the source codes using them. It was found that they completed the codes that can pass the test codes, whereas the use of exception handling functions was sometimes insufficient or incorrect.

The rest of this paper is organized as follows: Sections 2 and 3 introduce the TDD method and JPLAS respectively. Section 4 presents the test code generation method. Section 5 shows the evaluation result. Sections 6 shows related works. Finally, Section 7 concludes this paper with some future works.

2 Test-driven Development Method

In this section, we introduce the test-driven development method along with its features.

2.1 Outline of TDD Method

In the TDD method, the test code should be written before or while the source code is implemented, so that it can verify whether the current source code satisfies the required specifications during its development process. The basic cycle in the TDD method is as follows:

1) to write the test code to test each required specification,
2) to write the source code, and
3) to repeat modifications of the source code until it passes each test using the test code.

2.2 JUnit

In JPLAS, we adopt *JUnit* as an open-source Java framework to support the TDD method. *JUnit* can assist the unit test of a Java code unit or a *class*. Because *JUnit* has been designed with the Java-user friendly style, its use including the test code programming is less challenging for Java programmers. In *JUnit*, a test is performed by using a given method whose name starts from *assert*. This paper adopts the *assertThat* method to compare the execution result of the source code with its expected value.

2.3 Test Code

A test code should be written using libraries in *JUnit*. Here, by using the following **source code 1** for *MyMath* class, we explain how to write a test code. *MyMath* class returns the summation of two integer arguments.

source code 1

```
1  public class Math {
2    public int plus(int a, int b) {
3      return( a + b );
4    }
5  }
```

Then, the following **test code 1** can test the *plus* method in the *MyMath* class.

test code 1

```
1  import static org.junit.Assert.*;
2  import org.junit.Test;
3  public class MathTest {
4    @Test
5    public void testPlus() {
6      Math ma = new Math();
7      int result = ma.plus(1, 4);
8      assertThat(5, is(result));
9    }
10 }
```

The names in the test code should be related to those in the source code so that their correspondence becomes clear:

- The class name is given by the *test class name + Test*.
- The method name is given by the *test + test method name*.

The test code imports *JUnit* packages containing test methods at lines 1 and 2, and declares *MathTest* at line 3. *@Test* at line 4 indicates that the succeeding method represents the test method. Then, it describes the test method.

The test code performs the following functions:

1) to generate an instance for the *MyMath* class,
2) to call the method in the instance in 1) using the given arguments,
3) to compare the result with its expected value for the arguments in 2) using the *assertThat* method, where the first argument represents the expected value and the second one does the output data from the method in the source code under test.

2.4 Features in TDD Method

In the TDD method, the following features can be observed:

1. The test code can represent the specifications of the source code, because it must describe the function tested in the source code.
2. The test process for a source code becomes efficient, because each function can be tested individually.
3. The refactoring process of a source code becomes effective, because the modified code can be tested instantly.

Therefore, to study the TDD method and writing a test code is useful even for students, where the test code is equivalent to the source code specification. Besides, students should experience the software test that has become important in software companies.

3 Java Programming Learning Assistant System

In this section, we review the outline of our Java programming learning system *JPLAS*.

3.1 Server Platform

JPLAS is implemented as a Web application using *JSP/-Java*. For the server platform, it adopts the operating system *Linux*, the Web server *Apache*, the application server *Tomcat*, and the database system *MySQL*, as shown in Figure 1. For the browser, it assumes the use of *Firefox* with *HTML*, *CSS*, and *JavaScript*.

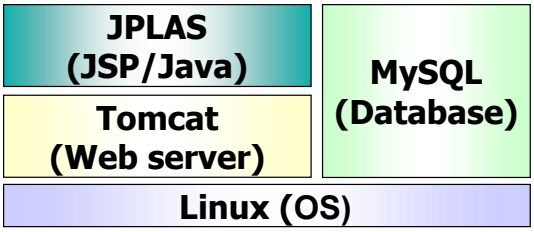


Figure 1: JPLAS server platform.

3.2 Teacher Service Functions

JPLAS has user functions both for teachers and students. *Teacher service functions* include the registration of courses, the registration and management of assignments, and the verification of source codes that are submitted by students. To register a new assignment, a teacher needs to input an assignment title, a problem statement, a reference (model) source code, and a test code. After the registration, they are disclosed to the students except for the source code. Note that the test code must be able to test the model code correctly. Using the correspondence between a source code and a test code in Section 2.3, it is possible to automatically generate a template for the test code from the source code. Then, a teacher merely needs to specify concrete values for the arguments in each test method to complete the test code.

To evaluate the difficulty of assignments and the comprehension of students, a teacher can refer to the number of submissions for code testing from each student. If a teacher finds an assignment with plenty of submissions, it can be considered as quite difficult for the students, and should be changed to an easier one. If a teacher finds a student who submitted codes in many times whereas other students did in a few times, this student may require additional assistance from the teacher.

3.3 Student Service Functions

Student service functions include the view of the assignments and the submission of source codes for the assignments. A student should write a source code for an assignment by referring the problem statement and the test code. It is requested to use the class/method names, the types, and the argument setting specified in the test code. JPLAS implements a Web-based source code editor called *CodePress* [11] so that a student can write codes on a Web browser. All submitted source codes will be stored in the database on the server as a reference for students.

4 Proposal of Test Code Generation Method

In this section, we propose the test code generation method for coding the standard input/output with exception handling.

4.1 Scope of Source Code under Test

At the early stage of the Java programming education, the responsibility of a student is to master how to write a source code that contains the standard input/output with exception handling. Thus, a teacher in a Java programming course should prepare a considerable number of assignments for writing source codes containing them, where many Java programming textbooks offer such assignments for novice students.

The source code in this paper must contain the functions for the standard input/output and the exception handling. Then, if the proper data is given to the code from the standard input, it must handle it correctly and outputs the message specified in the assignment to the standard output. On the other hand, if the improper data is given, it must handle it using the exception handling command without abortion and outputs the corresponding message.

4.2 Requirements in Test Code

Subsequently, the test code must satisfy the following requirements:

1. The input data from the standard input (keyboard) must be described in the test code to test the standard input in the source code.
2. The output data to the standard output (console) must be received by the test code to test the standard output in the source code.

3. The input data must be elaborated in the test code for the standard input.

4. The input data in the test code should cover any possible one for the standard input, including the proper and improper ones.

5. The expected output data for each input data must be narrated in the test code correctly.

4.3 Solutions for Requirements

Test code generation method adopts following functions and commands to solve the above mentioned requirements by referring the test code implementation in [10]:

- To describe the standard input data to the source code, the *Inputln* method in *StandardInputSnatcher* class is adopted in the test code. It is noted that *StandardInputSnatcher* class is extended from *InputStream* class.

- To receive the standard output data from the source code, the *readLine* method in *StandardOutputSnatcher* class is adopted in the test code. It is noted that *StandardOutputSnatcher* class is extended from *PrintStream* class.

- Any possible standard input data is prepared by a teacher beforehand. It is used in the argument of *Inputln*.

- To obtain the expected standard output data from the code for each input data, the reference source code is executed with this input data.

- Each pair of the standard input and output data is embedded into the test code.

4.4 Conditions of Source Code

Currently, to avoid the complexity, the proposed method confines the applicable source code that satisfies the following conditions:

1. it has the *main* method only.

2. it contains the standard input function.

3. it contains the standard output function for handling the proper input.

4. it contains the standard output function for handling the exception.

It is noted that a source code containing multiple standard input/output functions can be handled by increasing the number of *Inputln* or *assertThat* in the test code accordingly. Besides, if a code does not have the *main* method, it can be handled by describing the proper statements to execute the method for the standard input/output in the test code.

An example source code in this scope is as follows:

source code 2

```
1  import java.util.Scanner;
2  public class Sample {
3    public static void main(String args[]){
4      int number;
5      Scanner scan = new Scanner(System.in);
6      try{
7        System.out.print("Enter an integer");
8        String actual = scan.nextLine();
9        number = Integer.parseInt(actual);
10       System.out.println(number +": is input
         number");
11     } catch(NumberFormatException e) {
12       System.out.print("
         NumberFormatException occurs!");
13     }
14   }
15 }
```

source 2 accepts an integer data from a console and outputs a message with this data on a display. In this source code, 1) it has only the *main* method at line 3, 2) *scan* object of *Scanner* class is defined at line 5 as the standard input function, 3) *System.out.println* is called at line 10 as the standard output function for handling the proper input, and 4) *System.out.println* is called at line 12 as the standard output function for handling the exception.

4.5 Test Code Template

Then, the proposed method provides the *test code template* containing the required functions for the above mentioned source code. The following code describes the core part of the test code template starting from *@Test*. In advance, several *import* statements to use related libraries, and the instance generations for the *StandardInputSnatcher* and *StandardOutputSnatcher* classes are necessary. Besides, the definitions of these classes are also required to complete the test code template.

In this template, *in.Inputln* at line 29 gives the standard input data to the source code, where *in* is an instance of *StandardInputSnatcher* class. The statements at lines 30-37 run the source code and read the standard output data for this input data, where *out* is an instance of *StandardOutputSnatcher* class. *expected* at line 38 represents the expected output data of the source code. The blanks ″ ″ at lines 29 and 38 should be filled by the standard input and output data. *assertThat* at line 39 compares the expected data with the output data of the code. The whole statements at lines 25-40 should be prepared for each input data.

test code template

```
1  import static org.hamcrest.CoreMatchers.is;
2  import static org.junit.Assert.assertThat;
3  import static org.junit.Assert.*;
4  import java.io.InputStream;
5  import org.junit.Before;
6  import org.junit.Test;
7  import Snatcher.StandardOutputSnatcher;
8  import java.io.BufferedReader;
9  import java.io.ByteArrayOutputStream;
10 import java.io.IOException;
11 import java.io.InputStream;
12 import java.io.PrintStream;
13 import java.io.StringReader;
14
15 public class TemplateTest {
16   private StandardInputSnatcher in = new
       StandardInputSnatcher();
```

```
17      private StandardOutputSnatcher out = new
            StandardOutputSnatcher();
18
19      @Before
20      public void setUp() {
21        System.setIn(in);
22        System.setOut(out);
23      }
24
25      @Test
26      public void test1() throws Exception {
27        StringBuffer bf = new StringBuffer();
28        String actual,line,expected;
29        in.Inputln(""); // standard input
30        Sample.main(new String[0]);
31        System.out.flush();
32        while((line = out.readLine()) != null) {
33          if (bf.length() > 0)
34          bf.append("\n");
35          bf.append(line);
36        }
37        actual = bf.toString();
38        expected = ""; // expected standard output
39        assertThat(actual,is(expected));
40      }
41    }
```

4.6 Test Code Generation Procedure

The test code generation procedure using the *test code template* in the proposed method is as follows:

1) A teacher prepares the reference source code for the assignment.

2) He/she prepares a set of possible standard input data to the source code.

3) He/she runs the source code by using each standard input data and observes the corresponding standard output data.

4) He/she embeds the standard input data into ″ ″ at line 29 and the observed standard output data into ″ ″ at line 38 in the test code template.

As the possible standard input data in step 2), the following five data types should be considered. Then, the teacher needs to select one value for each data type, which is used in step 3).

- positive integer: 5
- negative integer: -14
- zero integer: 0
- floating-point number: 0.5
- one-byte character: abc
- two-byte character: A B C

4.7 Generated Test Code Example

This subsection introduces an example of the test code generated by applying the proposed method to **source code 2**. The file name for the generated test code is given as *SampleTest.java*. The following **test code 2** shows a part of the test code.

test code2

```
1   .....
2
3     @Test
4     public void test1() throws Exception {
5       StringBuffer bf = new StringBuffer();
6       String actual,line,expected;
7       in.Inputln("5"); // proper standard input data
8       Sample.main(new String[0]);
9       System.out.flush();
10      while((line = out.readLine()) != null) {
11        if (bf.length() > 0)
12        bf.append("\n");
13        bf.append(line);
14      }
15      actual = bf.toString();
16      expected = "Enter an integer" +
17       "5: is input number";
18      assertThat(actual,is(expected));
19    }
20
21    @Test
22    public void test2() throws Exception {
23      StringBuffer bf = new StringBuffer();
24      String actual,line,expected;
25      in.Inputln("abc"); // improper standard input
          data
26      Sample.main(new String[0]);
27      System.out.flush();
28      while((line = out.readLine()) != null) {
29        if (bf.length() > 0)
30        bf.append("\n");
31        bf.append(line);
32      }
33      actual = bf.toString();
34      expected = "Enter an integer" + "
          NumberFormatException occurs!";
35      assertThat(actual,is(expected));
36    }
37  .....
```

5 Evaluation

In this section, we evaluate the effectiveness of the proposed *test code generation method* in terms of generating test codes from existing source codes and writing source codes using the test codes by students.

5.1 Test Code Generation Results

First, we evaluate the method in generating test codes from source codes. For this purpose, 97 source codes were collected from Java programming textbooks or Web sites [12]-[16], and the test codes were generated by applying the proposed method. It is noted that some codes in [15] were modified to using the standard input/output through the console instead of using the dialog box. Then, the correctness of each test code was examined by testing the original source code. It was found that our method generated the test codes that can pass original codes correctly except for one source code, which outputs a random number generated in the code. Thus, the effectiveness of the proposed method was confirmed.

The following **source code 3** shows an example source code in [12] where the method successfully generates the test code shown in **test code 3**. It is noted that *try - catch* is used here instead of *throws* in the original source code.

source code 3

```
1  import java.io.*;
2  class Sample3 {
3    public static void main(String[] args) throws
       IOException {
4      try {
5        System.out.println("Enter two integers
         ");
6        BufferedReader br =
7          new BufferedReader(new
         InputStreamReader(System.in));
8        String str1 = br.readLine();
9        String str2 = br.readLine();
10       int num1 = Integer.parseInt(str1);
11       int num2 = Integer.parseInt(str2);
12       System.out.println("The sum is " + (
         num1+num2) + ".");
13     } catch(NumberFormatException e) {
14       System.out.print("
         NumberFormatException occurs!");
15     }
16   }
17 }
```

test code3

```
1  .....
2
3    @Test
4    public void test1() throws Exception {
5      StringBuffer bf = new StringBuffer();
6      String actual,line,expected;
7      in.Inputln("2"); in.Inputln("7");// proper
         standard input data
8      Sample.main(new String[0]);
9      System.out.flush();
10     while((line = out.readLine()) != null) {
11       if (bf.length() > 0)
12       bf.append("\n");
13       bf.append(line);
14     }
15     actual = bf.toString();
16     expected = "Enter two integers" +"The
         sum is 9.";
17     assertThat(actual,is(expected));
18   }
19
20   @Test
21   public void test2() throws Exception {
22     StringBuffer bf = new StringBuffer();
23     String actual,line,expected;
24     in.Inputln("0.5"); in.Inputln("-3");//
         improper standard input data
25     Sample.main(new String[0]);
26     System.out.flush();
27     while((line = out.readLine()) != null) {
28       if (bf.length() > 0)
29       bf.append("\n");
30       bf.append(line);
31     }
32     actual = bf.toString();
33     expected = "Enter two integers" + "
         NumberFormatException occurs!";
34     assertThat(actual,is(expected));
35   }
36 .....
```

5.2 Source Code Writing Results

Next, we evaluate the proposed method in writing source codes with generated test codes by five students who are currently studying Java programming and have same technical levels. For this purpose, we prepared the following three problems, where all the students completed the source codes that pass the test codes for any problem.

5.2.1 Problem #1

In problem #1, the code accepts an integer data from a console, and outputs a message with this data to a console, where **source 2** is the reference source code and **test 2** is the test code. The source code from a student is expected to use *NumberFormatException* to check the input data format. Then, three students use this class for the exception handling, and one uses *Exception*. However, one student does not use it where he implements the data format checking function.

5.2.2 Problem #2

In problem #2, the code accepts an integer index from a console, and outputs the indexed data from the data array. The student code is expected to use *ArrayIndexOutofBoundException* to check the range of the index. Then, only one student uses this class. The other students implement the index checking function in the codes. Two students use *IOException*, and two students do not use any class for the exception handling. No student use *NumberFormatException* to check the input data format, although the class was requested in problem #1. Unfortunately, many students cannot integrate the knowledge that has been studied sequentially.

5.2.3 Problem #3

In problem #3, the code accepts a file path from a console, and outputs the string at the first line in the file. The student code is expected to use *FileNotFoundException* or *IOException* to check the file path. Then, three students use *FileNotFoundException*, one uses *Exception*, and one uses *IOException*.

5.2.4 Summary of Student Applications

This simple experiment of our proposal shows that the students can generally complete source codes using standard input/output with exception handling that can pass the generated test codes. However, their use of the class for the exception handling is sometimes insufficient or incorrect. It has been observed that these students are not experts, which causes the difference in their source codes, although they have enough programming skills. To let them understand the correct use, it is necessary to improve the proposed method.

6 Related Works

In this section, we introduce some related works to this paper.

In [17], Fu presented a static exception-flow analysis that computes chains of semantically-related exception-flow links and reports entire exception propagation paths. These chains can be used, 1) to show the error handling architecture of a system, 2) to assess the vulnerability of a single component and the whole system, 3) to support the better testing of an error recovery code, and 4) to facilitate the tracing of the root cause of a logged problem.

In [18], Rashkovits showed that most of college students understand the concept of Java exception handling at the basic level, and the majority of them have difficulty in understanding advanced properties such as use of multiple exceptions, flow of control in the context of exceptions, handling exceptions further up the calling chain, catching and handling hierarchically related exceptions, and overriding methods that throw exceptions. They also provided a tutorial of exception handling, and quoted that exception handling is perceived as a relatively difficult task by novice programmers. In future works, we will consider to adopt their contributions.

In [19], Júnior presented a practical approach to preserve the *exception policy* in a system by automatically checking *exception handling design rules*. They are checked through executions of *JUnit* test cases with *dynamic mock objects* that are generated by the supporting tool. Four versions of *Mobile Media in SPL* were used to evaluate whether the policy was preserved or not. The results show that the approach can effectively detect violations on the policy of software product lines.

In [20], Nakshatri presented an empirical study of exception handling patterns in Java projects. It forces developers to think in sophisticated ways to handle the exceptions. In this study, empirical data was extracted from projects by analyzing data in *GitHub* and *SourceForge* repositories. The results were compared with recommendations for best practices in exception handling presented by Bloch [21]. It has been observed that most programmers ignore checking exceptions, and higher classes in the exception class hierarchy are more frequently used.

In [22], Brunet presented the concept of *design test* that automatically checks whether the code conforms to the specific *design rule* by using a test-like program. To support it, *DesignWizard (DW)* had been developed with a fully-fledged API that allows writing design tests for Java codes using *JUnit*. The proposal was applied to three software products in their group and student projects in the undergraduate course. The results showed that this approach was suitable to check conformance between the design rules and the code implementation. Moreover, it has been observed that both designers and programmers appreciated the design tests as executable documents that can be easily kept up to date.

In [23], Akahane presented a Web-based *automatic scoring system* for Java programming assignments to reduce loads of teachers in verifying a huge number of codes and in giving feedbacks to students. The system receives Java application programs submitted by students, and immediately returns the results of *JUnit* tests where the *Java Reflect API* is adopted for testing private classes and methods that have been commonly found in introductory courses. The regular expression is used to compare the output texts of each student program and those of the reference program. Through use in an actual course in their university, it was confirmed that this system was very helpful for students to improve programming skills by correcting mistakes in their programs and repeating their submissions.

In [24], Kitaya presented a Web-based scoring system of programming assignments to students, which is similar to JPLAS. Their test consists of compiler check, *JUnit* test, and result test. The result test verifies the correctness of a student code composed of only the *main* method that reads/writes data from/to the standard input/output devices, by comparing the results of this code and of the reference code. However, the method has several disadvantages from our proposal: 1) it is only applicable to a code composed of the *main* method with the standard input/output, 2) it uses other programs to use the redirection for handling the standard input/output, and 3) it needs several input files to check the correctness for different input data. On the other hand, our method is applicable to a code containing other than the *main* method, it needs only *JUnit* with a test code, and all the input data can be described in a single test code.

7 Conclusion

In this paper, we proposed the *test code generation method* for the code writing problem in JPLAS that requires implementing a Java source code containing the *standard input/output* with *exception handling*. To access the standard input/output from the test code on *JUnit*, the *test code template* is first prepared with the *input/output snatcher classes*. Then, the test code is completed by embedding the input and output extracted by running the *reference source code* into the template. This proposal is helpful in reducing the teacher load in writing the test code for the programming assignment that requires the standard input/output with exception handling, which is common for novice students. The effectiveness is evaluated through applying the method to 97 source codes in Java programming text books or Web sites, and asking five students to write source codes using the generated test codes for three problems. In future works, we will extend the proposed method to handle other input/output functions, other methods than the *main* method, and improve the readability of the generated test code to make it easier for novice students.

References

[1] N. Funabiki, R. Kusaka, N. Ishihara, and W.-C. Kao, "A proposal of test code generation tool for Java programming learning assistant system," Proc. IEEE Int. Conf. Adv. Inform. Netw. Appl., pp. 51-56, March 2017.

[2] S. Cass, The 2015 top ten programming languages, `http://spectrum.ieee.org/computing/software/the-2015-top-ten-programminglanguages/?utm_so.`

[3] N. Funabiki, Tana, K. K. Zaw, N. Ishihara, and W.-C. Kao, "A graph-based blank element selection algorithm for fill-in-blank problems in Java programming learning assistant system", IAENG Int. J. Comput. Science, vol. 44, no. 2, pp. 247-260, May 2017.

[4] N. Funabiki, Y. Matsushima, T. Nakanishi, K. Watanabe, and N. Amano, "A Java programming learning assistant system using test-driven development method", IAENG Int. J. Comput. Science, vol.40, no.1, pp. 38-46, Feb. 2013.

[5] K. K. Zaw, N. Funabiki, and W.-C. Kao, "A proposal of value trace problem for algorithm code reading in Java programming learning assistant system", Inf. Eng. Express, vol. 1, no. 3, pp. 9-18, Sep. 2015.

[6] N. Ishihara, N. Funabiki, and W.-C. Kao, "A proposal of statement fill-in-blank problem using program dependence graph in Java programming learning assistant system", Inf. Eng. Express, vol. 1, no. 3, pp. 19-28, Sep. 2015.

[7] N. Ishihara, N. Funabiki, M. Kuribayashi, and W.-C. Kao, "A software architecture for Java programming learning assistant system", Int. J. Comput. Soft. Eng., vol. 2, no. 1, Sep. 2017.

[8] K. Beck, Test-driven development: by example, Addison-Wesley, 2002.

[9] JUnit, http://www.junit.org/.

[10] Diary of kencoba, http://d.hatena.ne.jp/kencoba/20120831/1346398388.

[11] CodePress, http://codepress.sourceforge.net.

[12] M. Takahashi, Easy Java, 5th Ed., Soft Bank Creative, 2013.

[13] H. Yuuki, Java programming lessen, 3rd Ed., Soft Bank Creative, 2012.

[14] Y. D. Liang, Introduction to Java programming, 9th Ed., Pearson Education, 2014.

[15] Java programming seminar, http://java.it-manual.com/start/about.html.

[16] Kita Soft Koubo, http://kitako.tokyo/lib/JavaExercise.aspx.

[17] C. Fu and B. G. Ryder, "Exception-chain analysis: revealing exception handling architecture in Java server applications", Proc. Int. Conf. Soft. Eng., pp. 230-239, May 2007.

[18] R. Rashkovits and I. Lavy, "Students' understanding of advanced properties of Java exceptions", J. Inform. Tech. Edu., vol. 11, pp. 327-352, 2012.

[19] R. J. S. Júnior and R. Coelho, "Preserving the exception handling design rules in software product line context: a practical approach", Proc. Latin-American Symp. Depend. Comp. Work., pp. 9-16, 2011.

[20] S. Nakshatri, M. Hegde, and S. Thandra "Analysis of exception handling patterns in Java projects: an empirical study", Proc. IEEE/ACM Work. Conf. Mining Soft. Rep., pp. 500-503, May 2016.

[21] J. Bloch, Effective Java, 2nd Ed., Prentice Hall PTR, Upper Saddle River, NJ, USA, 2008.

[22] J. Brunet, D. Guerrero, and J. Figueredo, "Design tests: an approach to programmatically check your code against design rules", Proc. Int. Conf. Soft. Eng., pp. 255-258, May 2009.

[23] Y. Akahane, H. Kitaya, and U. Inoue, "Design and evaluation of automated Scoring Java programming assignments" Proc. Int. Conf. Soft. Eng., Art. Intel., Net. Para./Dist. Comp., pp.1-6, 2015.

[24] H. Kitaya and U. Inoue, "An online automated scoring system for Java programming assignments", Int. J. Inform. Edu. Tech., vol. 6, no. 4, pp. 275-279, April 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 38-48 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Framework for the Formal Specification and Verification of Security Guidelines

Zeineb Zhioua[*,1], Rabea Ameur-Boulifa[2], Yves Roudier[3]

[1]*EURECOM, Digital Security, France*

[2]*LTCI, Télécom ParisTech, Université Paris-Saclay, France*

[3]*I3S - CNRS - Université de Nice Sophia Antipolis, France*

ARTICLEINFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords:
Security Guidelines
Formal specification
Model Checking
Information Flow Analysis
Program Dependence Graph

ABSTRACT

Ensuring the compliance of developed software with general and application-specific security requirements is a challenging task due to the lack of automatic and formal means to lead this verification. In this paper, we present our approach that aims at integrating the formal specification and verification of security guidelines in early stages of the development lifecycle by combining both the model checking analysis together with information flow analysis. We present our framework that is based on an extension of LTS (labelled transition Systems) by data dependence information to cover the end-to-end specification and verification of security guidelines.

1 Introduction

This paper is an extension of work originally presented in Pacific Rim International Symposium on Dependable Computing (PRDC 2017) [1]. About 64% of the 2500+ vulnerabilities in the National Vulnerability Database NVD were due to programming mistakes [2], and the majority of software vulnerabilities are caused by coding errors. Flaws and errors can be introduced during the different phases of the software development lifecycle, from design to development. The missed programming errors can turn into security vulnerabilities at run-time, and can be exploited by intruders who may cause serious damage to the software critical assets and resources. The undetected flaws can cause a cost increase, comprising maintenance and flaw correction fees. Using code analysis tools would avoid such issues and help produce safe and secure software. The last decades have witnessed the development of many analysis techniques that aim at detecting security vulnerabilities in the early stages of development lifecycle; however, most attention was devoted to control-flow, somehow ignoring the data dependencies source of vulnerabilities that can remain undetected.

On the one hand, it is important for developers to be aware of domain-specific requirements as failure to pass the verification and validation phase, and subsequent application corrections and maintenance may be costly, time-consuming, and affect the company's reputation. On the other hand, security guidelines should be expressed in a way that allows their understanding and easy implementation for developers who may not be security experts to develop and deliver secure software.

Security guidelines or security best practices serve as recommendations to developers to reduce the application exposure to security issues, and to ensure that the developed system will behave as expected in hostile environments.

The problems described above entail the need to add on top of the development process different types of verification such as the compliance with the general and application-specific security requirements. Applying formal methods in the different phases of the software engineering process can help further understanding of the system, and detect design flaws rather early in the development.

In our papers [3,4] we conducted a survey on the static code analysis methods with the objective of identifying the approach that best fits our needs in terms of information flow properties detection and

[*]Corresponding Author: Zeineb Zhioua, zeineb.zhioua@eurecom.fr

https://dx.doi.org/10.25046/aj030106

validation, as well as on the system abstraction models that constitute a strong basis in order to carry out the analysis.

The main problem we are tackling in this paper is how to automatically verify the systematic application and compliance of (being) developed software with security requirements expressed in natural language. This problem requires the transformation of the guidelines written informally into a precise formalism by security expert(s), and this is a very tedious task. Nowadays, formal methods, in particular formal verification, are increasingly being used to enforce security and safety of programs.

We propose a framework that first provides security experts with the means to express security guidelines in a more formal way than plain text. Then, our framework verifies the adherence to the guidelines over an abstraction of the program, and provides understandable and clear feedback to the developer to indicate the exact program location where the error occurred. The innovation of our framework relies on the combination between model checking and data dependencies together with the analysis of the system behavior without actually executing it. We focus on the current version of our framework on the Java programming language.

Below we provide a sample code (Figure 1) that presents an implicit violation of the guideline from the *OWASP Cryptographic Storage Cheat Sheet* [4]:"Store unencrypted keys away from the encrypted data" [1] explaining the encountered risks when the encryption key is stored in the same location as the encrypted data. If we want to verify if the code below meets this guideline or not, then we have first to annotate the sources and the sinks, and run the analysis. We need first to highlight several elements in the codes below; the data key *k* and *encrypted_cc* are stored respectively in file **keys.txt** and **encrypted_cards.txt**. One may conclude that the guideline is met, as key *k* and *encrypted_cc* are stored in separate files. However, the two files are located in the same file system, which constitutes a violation of the guideline. Let us look at the details of the code. The developer encrypts the secret data credit card number, and stores the cipher text into a file. At line 115, the developer creates a byte array *y* used as parameter for the instantiation of a SecretKeySpec named *k* (line 116). At line 119, the key *k* is stored in a file, through the invocation of method *save_to_file* (Figure 1). Once created, key *k* is provided as parameter to the method *save_to_file(String data, String file)* (Figure 1) The developer then encrypts the secret variable creditCardNumber using method *private static byte[] encrypt(Key k, String text)* which uses key *k* as parameter. The encrypted data is then stored using method *save_to_file(String data, String file)* (Figure 1). One can conclude that the guideline is met, as key *k* and *encypted_cc* are stored in separate files. But if we take a closer look, we would notice that the two files are located in the same file system. Hence, the code violated the guideline.

In this paper, we go through the approach that we propose in order to help solving the difficulty of capturing implicit and subtle dependencies that can be source of security guidelines violations.

```
106  public static void main(String[] args)
107      throws NoSuchAlgorithmException,
108      NoSuchProviderException,
109      FileNotFoundException {
110  int c = 123456;
111  Payment p = new Payment();
112  p.setCreditCardNumber(c);
113
114  String x = "0xe04fd020ea3a6910a2d808002b30309d";
115  byte[] y = hexStringToByteArray(x);
116  SecretKeySpec k = new SecretKeySpec(y, "AES");
117
118  // save
119  save_to_file(k,"C:/[redacted]"
120      + "//[redacted]//src//secGuidelines//keys.txt");
121
122  // encrypted data
123  byte[] encrypted_cc = encrypt(k, Integer.toString
124      (p.getCreditCardNumber()));
125
126  // save
127  save_to_file(encrypted_cc,"C:/[redacted]"
128   + "//[redacted]//src//secGuidelines//encrypted_cards.txt");
129  }
```

Figure 1: Sample code for the encryption of credit card number

```
146  public static void save_to_file(String data, String file) {
147  try (PrintWriter out = new PrintWriter(file)) {
148      out.print(data +"\r\n");
149  } catch (FileNotFoundException e) {
150      // TODO Auto-generated catch block
151      e.printStackTrace();
152      System.out.println("file error");
153  }
154  }
```

Figure 2: Source code of save_to_file method

The flow of this paper is as follows: Section 2 introduces the security guidelines and discusses the motivation behind this work. In Section 3, we explain in detail the approach we carried out. We highlight several issues related with the guidelines presentation to developers in Section 4. In Section 5, we introduce the notion of information flow analysis that we make use of in our framework for the detection of implicit dependencies. In Section 6, we present our model construction methodology. We provide in Section 7 a comprehensive application of our formalism to specify and to verify the selected guideline. In section 8, we outline the related work, followed by a discussion in Section 9. And finally, Section 10 concludes the paper and discusses future work.

[1] https://www.owasp.org/index.php/Cryptographic_Storage_Cheat_Sheet#Rule_-_Store_unencrypted_keys_away_from_the_encrypted_data

2 Security Guidelines

Organizations and companies define non-functional security requirements to be applied by software developers, and those requirements are generally abstract and high-level. Security requirements such as confidentiality and integrity are abstract, and their application requires defining explicit guidelines to be followed in order to meet the requirements. Security guidelines describe bad as well as good programming practices that can provide guidance and support to the developer in ensuring the quality of his developed software with respect to the security aspect, and hence, to reduce the program exposure to vulnerabilities when delivered and running on the customer platform (on premise or in the cloud). Bad programming practices define the negative code patterns to be avoided, and that can lead to exploitable vulnerabilities, while good programming practices represent the recommended code patterns to be applied on the code.

Official sources, such as OWASP [5], Oracle [6], CERT [7], NSA [8], NIST [9] propose rules and examples of good/bad programming practices. The presentation of the security guidelines differ from one source to another. For instance, CERT Oracle Coding Standard for Java [7] provides for each guideline a textual description, followed by a compliant sample code, and another sample code violating the guideline. OWASP [5] provides for most guidelines a detailed description, and examples of compliant and non-compliant solutions.

Motivation The OWASP Foundation[5] for instance introduces a set of guidelines and rules to be followed in order to protect data at rest. However, the guidelines are presented in an informal style, and their interpretation and implementation require security expertise, as stressed in [3]. In the *OWASP Storage Cheat Sheet* [5], OWASP introduces the guideline "Store unencrypted keys away from the encrypted data" [5] explaining the encountered risks when the encryption key is stored in the same location as encrypted data. This guideline recommends not to store encrypted together with the encryption key, as this operation can result in a compromise for both the sensitive data and the encryption keys. However, encryption keys can be declared as byte arrays with insignificant names, which makes their identification as secret and sensitive data very difficult.

Correctly applying this guideline would provide a strong protection mechanism against this attack scenario: an attacker can get access to the encryption server or client, and can retrieve the encrypted data with the encryption key. Fetching those two elements allows the deciphering of the encrypted sensitive information. This reminds the well known HeartBleed[2] [10] attack that occurred couple of years ago (April 2014), and that allowed an attacker to read the memory, steal users credentials directly from the systems protected by the vulnerable version of OpenSSL. This example emphasizes the critical attacks that can be performed if the guideline is not respected. OWASP provides a set of security guidelines that should be met by developers, but does not provide the means to ensure their correct implementation. We aim at covering this gap through the formal specification of security guidelines and their formal verification using formal proofs.

In the sample code that we provided in Section 1, we showed how we could implicitly violate the guideline. Detecting this violation is not trivial, as it includes subtle dependencies that should be analyzed with due consideration. This is the main objective of this paper, and we could achieve this through the approach that we depict in details in the next Section.

3 Approach

In this section, we go through the details of our approach that aims at filling the gap between the informal description of security guidelines presented in natural language, and their automatic verification on the code level to provide precise and comprehensive feedback to the developer.

We started first by doing attempts to extract security properties from the code level, but we found out that we needed to have a reference against which we can compare the extracted program parts. This brought the idea of performing a deep survey on the guidelines that we gathered from different sources, as explained in Section 4.

The positive (resp. negative) security guidelines serve to express the desired (resp. undesired) program behavior. However, we operate on the code level, meaning that we do not monitor the program execution. We need then to approximate the program behavior but still from a static point of view. This requires that we transform the program into a formal model that allows us to exploit its properties, and approximate its behavior. In addition, the program model that we construct should be able to represent the whole flow of information in order to be able to reason about how data propagates, and capture possible information leakage. This induces the need to choose a formalism to represent the information flows that can be checked over this program model. Having covered the aforementioned aspects, we proceed to verify whether the program meets the specified guidelines or not. As we aim to decrease considerably the heavy load on the developer, we propose an automatic formal verification of the guidelines. The purpose is to verify that security guidelines are met, and not to prove that the program is correct. To the best of our knowledge, no prior work has filled in this gap between the informal description of security guidelines and their automatic formal verification on the code level.

[2] http://heartbleed.com/

The big picture of the proposed approach is shown in Figure 3, highlighting the relevant steps towards fulfilling the transformation of security guidelines written in natural language into exploitable formulas that can be automatically verified over the program to analyze.

The crucial part of the work is the explicit mapping between abstract security guidelines formal specification, and concrete statements on the code.

In order to make the process more concrete, a *separation of duties* needs to be made. We make the distinction between the *security expert* and the *developer*. The former carries out the formal specification of the security guidelines and their translation from natural language to formulas and patterns. He establishes also the mapping between the abstract labels and possible Java language instructions. The latter invokes the framework that makes use of this specification to make the mapping between abstract labels and the program logic, and then to verify the compliance of his developed software with the security requirement.

The idea we propose, as depicted in Figure 3, is the following:

3.1 Formal Specification of Security Guidelines

Starting from the guidelines presented in an informal manner, we make the strong assumption that the **security expert** formally specifies the security guidelines by extracting the key elements, and builds the formulas or patterns based on formalism. The established formulas or patterns can be supported by standard model checking tools. We present in Section 7 how the guideline that we consider in this paper can be modeled in a formalism, and can be formally verified using a model checking tool.

3.2 Program Model Construction

Choosing a program representation depends on the intended application. In our case, the program should be abstracted in a way that preserves its properties, such as the explicit and the implicit dependencies, hence allowing the performance of deep information flow analysis. In our framework, we have chosen the Program Dependence Graph (PDG) (see section 6) as the representation model, for its ability to represent both control and data dependencies. The generated PDG is then augmented with details and information extracted from the formulas and patterns of the security guidelines. We performed Information Flow Analysis over the constructed PDG in order to augment it with further security-related details. This analysis aims at capturing the different dependencies that may occur between the different PDG nodes, hence, augmenting the generated PDG with relevant details, such as annotations mapping the PDG nodes to abstract labels of the security guidelines. Then, we generate form the augmented PDG, a Labelled Transition System tha is accepted by model checking tools.

3.3 Verification

As previously mentioned, security guidelines will be modeled in the form of sequence of atomic propositions or statements representing the behavior of the system. The security guidelines will then be verified over the Labelled Transition System that we generate from the PDG augmented with implicit and subtle dependencies. The verification phase can have the following outcomes:

- The security guideline is valid over all the feasible paths
- The security guideline is violated

The first case can be advanced further, meaning that the verification can provide more details to the developer (or the tester) about circumstances under which the security guideline is valid. In the second case, recommendations to make the necessary corrections on the program can then be proposed . The concrete mapping between the abstract propositions in the formal security guidelines and the program model is managed in the Security Knowledge Base (Section 9).

4 Security Guidelines Analysis

Security guidelines are usually presented in an informal and unstructured way. Their presentation differs from one source to another, which can be misleading to developers. We did the effort of analyzing the guidelines from different sources, and we raised different problems that we have discussed in [4]. Upon this survey, we noticed a lack of precision and a total absence of automation. From developers perspective, the understanding and interpretation of guidelines is not a trivial task, as there is no formalization that exposes the necessary program instructions for each guideline, or that explains how to apply them correctly in their software. In order to overcome this weakness, we come up with a centralized database that gathers the possible mappings between guidelines and Java instructions in the Security Knowledge Base. In the OWASP Secure Coding Practices guide [5], a set of security guidelines are presented in a checklist format arranged into classes, like Database Security, Communication Security, etc. The listed programming practices are general, in a sense that they are not tied to a specific programming language. Another programming practices guide we can consider for instance is the CERT Oracle Coding Standard for Java[7] ; for each guideline, the authors provide a detailed textual explanation. For most, there are also provided examples of compliant and non-compliant sample codes in addition to the description. We want to pinpoint another key element that gathered our attention; there is a huge effort invested in order to build and maintain the catalogs, but no attempt was undertaken to instrument their automatic verification on the code level.

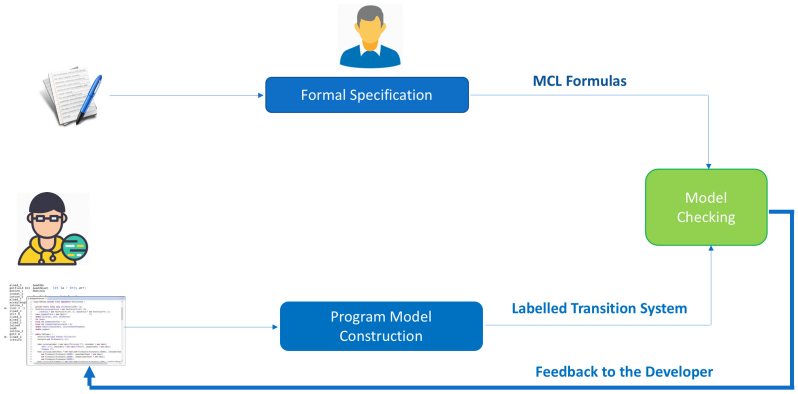


Figure 3: Framework for the formal specification and verification of Security Guidelines

5 Information Flow Analysis

In our framework, we make use of the Information Flow Analysis for many purposes, mainly the detection of implicit and subtle dependencies that can be source of covert channels and sensitive information leakage. From a security perspective, this might have serious damages on the security of the infrastructure as well as on users sensitive data. One might argue about using known security mechanisms such as access control to control the information propagation in a program. This aspect is of a paramount importance when dealing with information security. However, from a historic point of view, access control mechanisms, are used to verify the access rights at the point of access, and then, to allow or deny the access to the asset over which the mechanism is set. Access control mechanisms, just like encryption, can't provide assurance about where and how the data will propagate, where it will be stored, or where it will be sent or processed. This entails the need for controlling information flow using static code analysis. This same idea is emphasized by Andrei Sabelfeld and Andrew C. Myers [11], who deem necessary to analyze how the information flows through the program. According to the authors, a system is deemed to be secure regarding the property confidentiality, if the system as a whole ensures this property.

The main objectives of the information flow control [12] are to preserve the confidentiality and integrity of data; the former objective consists in guaranteeing that confidential data don't leak to public variables. As for the second objective, it consists in verifying that critical data is independent from public variables/output. Information flow control analyzes the software with the objective of verifying its compliance and conformance to some security policies. Different approaches have been proposed for Information Flow Control, where we can distinguish between *language-based* and *type-based* information flow control. The former has the advantage of exploiting the program source code and the programming language specificity, but falls short in covering different aspects such as physical side channels that is covered by other approaches[13] and execution environment properties. Language-based security mechanisms have been treated in the literature, including the bytecode verifiers and sandbox model. Those mechanisms enforce security through the Java language, but only the bytecode verifiers make use of the static code analysis. Type-based information flow control, on the other hand, basically makes use of the typing rules that capture illegal flows of information throughout a program, however, they are neither flow-insensitive, context-sensitive nor object-sensitive, which leads to imprecision, which in turn leads to false alarms.

6 Program Model Construction

The starting key element for this step is the standard PDG that we generate from the Java program bytecode using the JOANA tool [14]. In this PDG, control and (explicit/implicit) data dependencies are captured, which constitutes a strong basis to perform a precise analysis. Since our main objective is to automatically verify the adherence of programs to formalized security guidelines, we need to model check the guidelines MCL formulas over the program model. However, the PDG is not formal, and doesn't consist a basis for the formal verification through model checking. Thus, we need to construct from the augmented PDG a model that is accepted by a model checking tool, and that can be verified automatically through model checking techniques. We depict in Figure 4 the program model construction flow that we have adopted to generate the Labelled Transition System (LTS) from the PDG, that we generate from the program sources.

- Augmented Program Dependence Graph: this component first builds the program dependence graph (PDG) from the Java bytecode (.class) using the JOANA IFC tool [14]. We have chosen the Program Dependence Graph (PDG) as the abstraction model for its ability to represent both control and (explicit/implicit) data dependencies. The generated PDG is then annotated

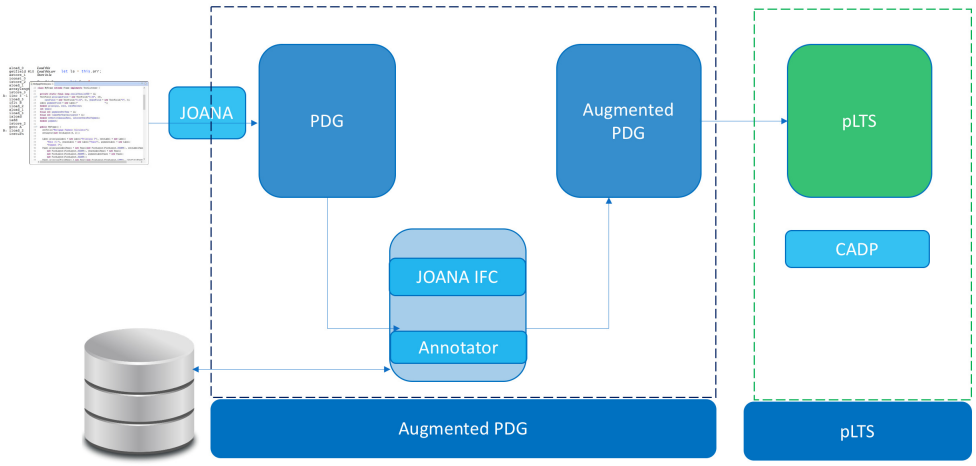


Figure 4: Methodology for the model construction: From program sources to the Augmented Program Dependence Graph, to the Labelled Transition System

by the **PDG Annotator** with specific annotations (labels in the MCL formulas). The **PDG Annotator** retrieves the nodes details (method signature) from the PDG, and fetches from the Security Knowledge Base the matching label if it exists. We run the information flow analysis using the JOANA IFC, that is formally proven [14] in order to capture the explicit and the implicit dependencies that may occur between the program variables. The operation results in a new PDG that we name the **Augmented PDG**. We show in Figure 5 the Augmented PDG of the sample code (Figure 1) that we consider in this paper.

- LTS Construction: this component translates automatically the **Augmented PDG** into a parameterized Labelled Transition System (pLTS) that is accepted by model checking tools. The annotations on the PDG nodes are transformed into labels on the transitions in the pLTS.

- Java Classes Parser: This component that we have developed [3] takes as input the URL of the Java class official documentation, and parses the HTML code (Javadoc) in order to extract all the relevant details: the class name, the inheritance, the description, the attributes, the constructor(s), the methods signatures, their return type and their parameters. This component populates the Security Knowledge Base with the extracted information.

6.1 Program Dependence Graphs

PDG (Program Dependence Graph) is a language-independent representation of program. This model was first proposed by Ferrante et al. [15] as a program representation taking into consideration both control and data relationships in a program. Formally, the PDG is a directed graph whose nodes correspond to program statements and whose edges model dependencies in the program. Those dependencies can be classified as either control or data dependencies. The nodes are predicates (variable declarations, assignments, control predicates) and edges are data and control dependencie representation; both types are computed using respectively control-flow and data-flow analysis.

PDGs have the ability to represent the information flow in a program, and have different properties, such as being *flow-sensitive*, *context-sensitive* and *object-sensitive* [13]. Being flow-sensitive is the ability of considering the order of statements in the program. The context-sensitivity is perceived from the fact that if the same method is invoked multiple times, then each call site will be represented a separate node in the graph, and will be analyzed separately. In other words, the methods calling context is considered, and this increases precision. The object sensitivity, on the other hand, is the ability to extend the analysis to the attributes level for Object Oriented Programs; object, which is an instance of a class, is not considered as an atomic entity, hence the analysis will be extended to the attributes level.
PDG abstracts away irrelevant details, such as independent and non-interacting program statements, that represent the unfeasible paths.

In our framework, we make use of the JOANA tool, that stands for Java Object sensitive Analysis [16] for the construction of the PDG. JOANA [14] is a framework that statically analyzes the byte code of Java programs; the tool first generates from the program byte code a PDG, which constitutes an over-approximation of the information flow in the program to analyze. The PDG contains apart from the nodes representing the statements and the variable declarations in the

[3] https://github.com/zeineb/Java-classes-parser

program, contains also edges referring to control and data dependencies between nodes. The dependencies represent explicit dependencies as well as transitive and implicit dependencies.

JOANA's strength relies on the ability to track how information propagates through a program, and captures both the explicit and implicit information flows.

6.2 Augmented Program Dependence Graph

We extend the definition of program dependence graphs to accommodate propositions over the set of program variables. Once the PDG is built, we compute all the propositions that are defined over the program. These propositions are parametrized by the variables defined within the program. Each node is annotated with a set of propositions.

There are two categories of information which can all be used as annotations. The first category is obtained by identifying the set of standard instructions: variables assignments, method calls, etc. at a given node. Example of methods are *encypt*, *hash*, *log*, *normalize*, *sanitize*. Second category is dedicated to relationship between variables, the dependencies in terms of explicit and implicit data dependencies between variables in a program.

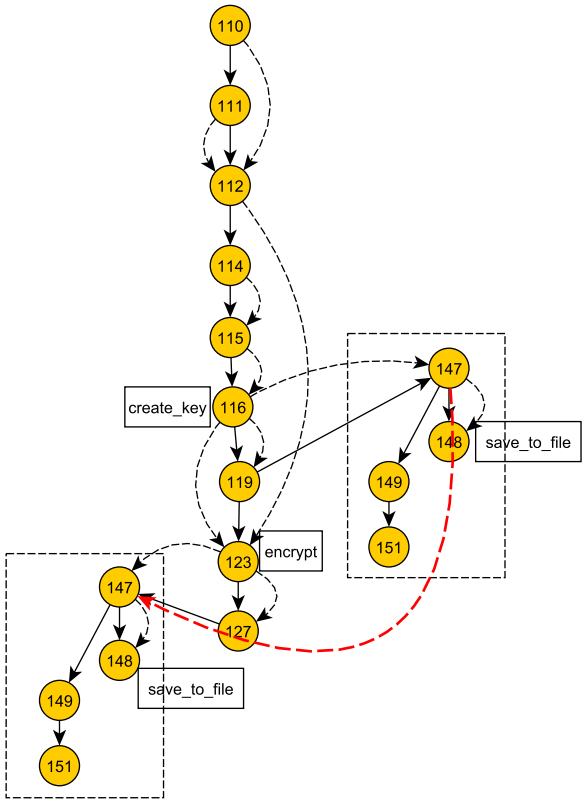


Figure 5: Augmented Program Dependence Graph for the sample code. Strong edges represent the control flows, the dashed edges refer to explicit and implicit data flows. Nodes are labeled with their corresponding instructions line numbers

Note that referring to our security knowledge base, different annotations on the PDG were pre-computed, like for instance the *save*, *userInput* and the *encyption_key*.

Automatic annotations. First, we have created our own annotations based on the atomic propositions of the security guidelines formulas. We made modifications on the source code of JOANA, and added the annotations *hash*, *userInput*, *Password*, *encrypt*, etc. in addition to the predefined annotations *SOURCE* and *SINK*. As shown in Figure 5, different program nodes are annotated with abstract labels, such as node 65 annotated as *Password*, node 73 annotated as *hash* and node 80 annotated as *log* or *store*. The log annotation was pre-computed after the PDG is built, meaning that the security knowledge base was accessed to fetch the concrete possible mappings between known APIs, methods, methods parameters mapped to the abstract labels the formulas are built upon. The mapping between the method invocation *logger.log* and the label *log* is already established. Same for the *hash* label. However, for the *Password*, the automatic annotation requires a semantic analysis to be performed over the code in order to determine the variable names matching password. The semantic analysis is not in the scope of this paper.

Annotations validation by the developer. Once the automatic detection of the atomic propositions on the PDG is performed, the intervention of developer is required to validate the added annotations. There might also be the case where the developer creates a method implementing the hash functionality, then the detection of the label *hash* on the program model will fail. In the sample code, the logging, which is one simple possible storing operation, was invoked. The developer, in our example, annotated the node 65 as *Password*, and the node 80 as *store* (in addition to the *log* automatic annotation).

6.3 Labelled Transition System

A parameterized Labelled Transition System (pLTS) is a labelled transition system with variables; a pLTS can have guards and assignment of variables on transitions. Variables can be manipulated, defined, or accessed in states, actions, guards, and assignments. JML [17], Z [18], B [19] allow to describe the states of the system through mathematics-based objects (machines, sets, etc.), and they describe pre- and post-conditions on the transitions between the states. Those languages deal with sequential programs and do not handle value passing for most.

Definition 1 (pLTS) *A parametrized LTS is a tuple* $pLTS \triangleq (S, s_0, L, \rightarrow)$ *where:*

- S *is a set of states.*
- $s_0 \in S$ *is the initial state.*

- *L is the set of labels of the form* $\langle \alpha, e_b, (x_j := e_j)^{j \in J} \rangle$, *where* α *is a parametrized action,* e_b *is a guard, and the variables* x_j *are assigned the expressions* e_j. *Variables in are assigned by the action, other variables can be assigned by the additional assignments.*
- $\rightarrow \subseteq S \times L \times S$ *is the transition relation.*

Informally, we interpret the behavior of a program as a set of reachable states and actions (instructions) that trigger a change of state. The states express the possible values of the program counter, they indicate whether a state is an entry point of a method (initial state), a sequence state (representing standard sequential instruction, including branching), a call to another method, a reply point to a method call, or a state which is of the method terminates. Each transition describes the execution of a given instruction, so the labels represent the instruction names.

The LTS labels can mainly be of three types: actions, data and dependencies.

- Actions: they refer mainly to all program instructions, representing standard sequential instructions, including branching and method invocations.
- Value passing: as performed analysis involves data, generated LTSs are parametrized, i.e, transitions are labelled by actions containing data values.
- Dependencies: in addition to program instructions, we added transitions that bring (implicit and explicit) data dependencies between two statements with the objective of tracking data flows. Indeed, transitions on LTS show the dependencies between the variables in the code. We label this kind of transition by *depend var1 var2* where *var1* and *var2* are two dependent variables.

7 Verification

With the objective of achieving our main goal consisting in helping a programmer verify that his program satisfies given security guidelines, we translate the augmented PDG into a formal description, which is precise in meaning and amenable to formal analysis. As usual, in the setting of distributed and concurrent applications, we provide behavioral semantics of analyzed programs in terms of a set of interacting finite state machines, called LTS [20]. An LTS is a structure consisting of states with transitions, labeled with actions between them. The states model the system states; the labeled transitions model the actions that a system can perform. Considered LTS are specific; their actions have a rich structure, for they take care of value passing actions and of assignment of state variables. They encode in a natural way the standard instructions of PDGs (as shown in Figure 5). Besides the classical behavior of a PDG, we encode in our LTS the result of tracking of explicit and implicit dependencies between program instructions. These dependencies are encoded by transitions labeled with the action *depend input_data output_data* (see Figure 6), allowing one to prove information flow properties.

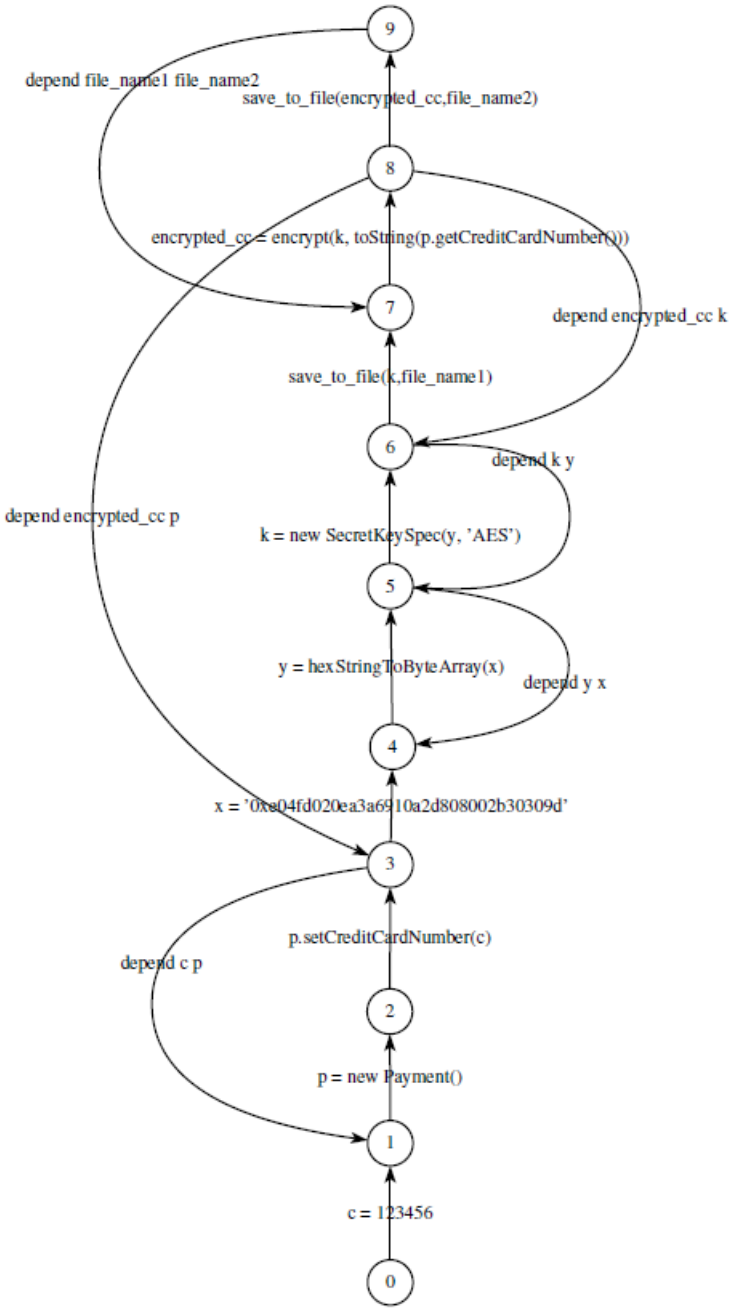


Figure 6: Labelled Transition System for the sample code given in Figure 1

Once the behavioral models are generated, we use model checking technique to automatically verify correctness of guidelines against the model.

For expressing the properties, we adopt MCL logic [21]. MCL (Model Checking Language) is an extension of of the alternation-free regular μ-calculus with facilities for manipulating data in a manner consistent with their usage in the system definition. The

MCL formula are logical formula built over regular expressions using boolean operators, modalities operators (necessity operator denoted by [ ] and the possibility operator denoted by ⟨ ⟩) and maximal fixed point operator (denoted by μ). For instance, the guideline *"Store unencrypted keys away from the encrypted data"* will be encoded directly by the following formula MCL:

```
[true*.{create_key ?key:String}.true*.
({save !key ?loc1:String}.true*.
{encrypt ?data:String !key}.true*.
{save !data ?loc2:String}.true*.
{depend !loc1 !loc2}
|
{encrypt ?data:String !key}.true*.
{save !key ?loc1:String}.true*.
{save !data ?loc2:String}.true*
.{depend !loc1 !loc2})] false
```

This formula presents five actions: the action {*create_key ?key:String*} denoting encryption key *key* (of type *String*) is created, the actions {*save !key ?loc1:String*}, {*save !data ?loc2:String*}, {*encrypt ?data:String !key*} denoting respectively the storage of the corresponding *key* in location *loc1*, the storage of the corresponding *data* in location *loc2*, the encryption of *data* using *key*, and the particular action *true* denoting any arbitrary action. Note that actions involving data variables are enclosed in braces ({ }). Another particular action that we make use of in this formula is {*depend !loc1 !loc2*}, denoting the implicit dependency between the file locations *loc1* and *loc2*; we captured this implicit dependency through advanced information flow analysis on the code.

This formula means that for all execution traces, undesirable behavior never occurs (false). The unexpected behavior is expressed by this sequence of actions: if encryption key *k* is saved in *loc1*, and *k* is used to encrypt *data* that is afterwords stored in *loc2*, then if *loc1* and *loc2* are dependent, the guideline is violated. The second undesirable behavior, expressed in the second sequence of the formula, means that if encryption of data using *k* occurs before the storage of *k* in *loc1*, and if *loc1* and *loc2* are dependent, then the guideline is violated.

We made use of the checker EVALUATOR of the CADP toolbox [22] to verify the property. From a behavioral point of view, the verification result is true, indicating that the guideline is verified. However, from a security point of view the answer should be false, as the variable *xx* containing the password in plain text, was leaked to the logging operation through and the implicit flow between this variable and the logging operation. To guarantee the reliability of the analysis, one needs to check secret/sensitive variables and depending variables as in the presented formula.

No surprise the answer is false. In addition to a *false*, the model checker produces a trace illustrating the violation from the initial state, as in Figure 7.

8 Related Work

Prior work in the area of information-flow security [23] has been developed during the last decades. A line of work [24] [25] adopts the Extended Static Checking, a specific technique for finding source code errors at compile-time. Eau Claire [21] framework operates as follows; it translates C program into Guard Commands (Guarded Command Language), that are afterwards translated into verification conditions for each function of the program. The generated verification conditions serve as input to automatic theorem prover. Adopting the prototype Eau Claire is very much-time consuming, requiring annotations entered by the developer, hence it is hard to integrate in the development phase.

De Francesco et al. [26] combine abstract interpretation and model checking to check secure information flow in concurrent systems. The authors make use of the abstract interpretation to build a finite representation of the program behavior: a labelled transition system . The security properties are specified in temporal logics, and are model checked over the built LTS. Their approach consists in verifying the non-interference property, meaning that the initial values of high level (secret) variables do not influence the final values of low level (public) variables. This approach checks for the non-interference only on two program states, which might miss possible information leakage within the process itself. In addition, the adopted formalism does not support value passing, and does not reason about data propagation.

The Verification Support Environment [27] is a tool for the formal specification and verification of complex systems. The approach adopted by the authors is similar to model-driven engineering, in the sense that the formal specification results in code generation from the model.

SecureDIS [28] makes use of model checking together with theorem-proving to verify and generate the proofs. The authors adopt the Event-B method, an extension of the B-Method, to specify the system and the security policies. The authors do not make it clear how the policies parameters are mapped to the system assets, and they do not extend the policy verification and enforcement on the program level. The work targets one specific system type (Data Integration System), and is more focused on access control enforcement policies, specifying the subject, the permissions and the object of the policy. However, access control mechanisms are not sufficient for the confidentiality property, as they can't provide assurance about where and how the data will propagate, where it will be stored, or where it will be sent or processed. The authors target system designers rather than developers or testers, and consider a specific category of policies focused on data leakage only.

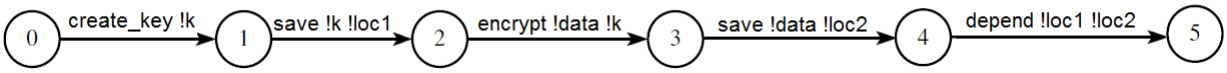


Figure 7: Path violating the guideline

GraphMatch [29] is a code analysis tool/prototype for security policy violation detection. GraphMatch considers examples of security properties covering both positive and negative ones, that meet good and bad programming practices. GraphMatch is more focused on control-flow security properties and mainly on the order and sequence of instructions, based on the mapping with security patterns. However, it doesn't seem to consider implicit information flows that can be the source of back-doors and secret variables leakage.

PIDGIN [30] introduces an approach similar to our work. The authors propose the use of PDGs to help developers verify security guidelines throughout the exploration of information flows in their developed software and also the specification and verification of adherence to those policies. Privacy policies are encoded in LEGALEASE language that allows to specify constraints on how user data can be handled, through the clauses ACCEPT and DENY [**?**]. The specification and verification of security properties rely on a custom PDG query language that serves to express the policies and to explore the PDG and verify satisfiability of the policies. The parameters of the queries are labels of PDG, which supposes that the developer is fully aware of the complex structure of PDGs, identify the sensitive information and the possible sinks they might leak to. For example, the authors propose a policy specifying that the guessing game program should not choose a random value that is deliberately different from the user's guess provided as input.

9 Discussion

We tackled in this paper the problem of verifying the adherence of the developed software to security guidelines that are presented in formal language. We raised the main issue regarding the interpretation, implementation and verification of the guidelines that are written in natural language, which might be subject to misinterpretation by developers. We worked towards stripping away ambiguities in [19] capturing implicit information flows that can be source of information leakage. We provided a centralized repository (Security Knowledge Base) gathering the security guidelines and patterns that we have discussed in detail in [19]. Apart from the patterns, we have centralized the labels that are used to build the formulas, mapped to the traditional information annotations (*SOURCE*, *SINK* and *DECLASS*) together with their security level (*HIGH* / *LOW*).

Security Knowledge Base *Security Knowledge Base* is a centralized repository gathering the labels of the formulas mapped to APIs, instructions, libraries or programs. This helps the automatic detection of labels on the system model. We built **Security Knowledge Base** using a Java classes parser [4] that operates as follows: for the different Java classes used in the program to analyze, we launch the parsing of this given class (html code, javadoc), and we extract all the relevant details, such as the description, the attributes, the constructors, the methods signatures and their parameters. Then, we made the effort of performing a semi-automatic semantic analysis to detect key elements, such as the keyword *secure*, *key*, *print*, *input*, etc. This operation is of a paramount importance, as it allows us to map the key words used to build the formulas, to the possible Java language instructions (methods invocations, constructors invocations, specific data types declarations, etc). For example, the Java API *KeyGenerator.generateKey()* is mapped to the label *isKey*. This label is also mapped to the traditional information flow annotation **high** level **source**.

As part of the Security Knowledge Base, we have also considered the vulnerabilities the program could eventually be exposed to if the guideline is violated. The Security Knowledge Base is rich yet extensible repository, that can be extended if new security concepts are introduced. For instance, the same guideline might be expressed through different MCL formulas and using different terms that are semantically equivalent. Let us take the example of the guideline *"Store unencrypted keys away from the encrypted data"* that we have previously formulated using those keywords: *create_key*, *save* and *encrypt*. Among the keywords contained in the dictionary that we provide to the security expert, the *create_key* is semantically equivalent to *isKey*, the word *save* is equivalent to *store*, and so on and so forth. Hence, the used keywords can be replaced with their equivalent as long as they do not alter the semantics of the guideline.

10 Conclusion

In this paper, we presented on a high level the approach that we propose with the objective of filling the gap for the verification of the security guidelines. We pinpoint different issues with the security guidelines that are present in different sources, but no verification means is provided to the developers to make sure that their being developed software adheres to those guidelines. Security guidelines are meant mainly for developers, but the way they are

[4] https://github.com/zeineb/Java-classes-parser

presented presents ambiguities, and this might lead to misinterpretation. Formalizing the guidelines would help strip away ambiguities, and prepare the ground for the formal verification. We stressed on the need for performing model checking as verification approach. This allows to have an automatic verification, hence to reduce the intervention of a human operator, whether the developer or the security expert leads this verification.

References

1. Zeineb Zhioua and Stuart Short and Yves Roudier, "Towards the Verification and Validation of Software Security Properties Using Static Code Analysis", in International Journal of Computer Science: Theory and Application,
2. Jon Heffley, Pascal Meunier, "Can Source Code Auditing Software Identify Common Vulnerabilities and Be Used to Evaluate Software Security?", in Proceedings of the 37th Hawaii International Conference on System Sciences-2004, https://doi.org/10.1109/HICSS.2004.1265654
3. Zeineb Zhioua and Stuart Short and Yves Roudier, "Static Code Analysis for Software Security Verification: Problems and Approaches", in 2014 IEEE 38th Annual International Computers, Software and Applications Conference Workshops,
4. Zeineb Zhioua and Yves Roudier and Stuart Short and Rabea Ameur-Boulifa, "Security Guidelines: Requirements Engineering for Verifying Code Quality", in ESPRE 2016, 3rd International Workshop on Evolving Security and Privacy Requirements Engineering
5. OWASP, Cryptographic Storage Cheat Sheet
6. Oracle, Secure Coding Guidelines for Java SE
7. CERT, SEI CERT Oracle Coding Standard for Java
8. NSA, Juliet Test Suite
9. National Institute of Standards and Technologies and Elaine Barker, "Recommendation for Key Management", 2016, http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-57pt1r4.pdf
10. Durumeric, Zakir and Kasten, James and Adrian, David and Halderman, J Alex and Bailey, Michael and Li, Frank and Weaver, Nicolas and Amann, Johanna and Beekman, Jethro and Payer, Mathias and others, "The matter of heartbleed", in Proceedings of the 2014 Conference on Internet Measurement Conference
11. Sabelfeld, Andrei and Sands, David, "Declassification: Dimensions and Principles"
12. Denning, Dorothy E. and Denning, Peter J, "Certification of Programs for Secure Information Flow", https://doi.org/10.1145/359636.359712
13. Hammer, Christian and Krinke, Jens and Snelting, Gregor, "Information flow control for java based on path conditions in dependence graphs" in IEEE International Symposium on Secure Software Engineering
14. Jrgen Graf and Martin Hecker and Martin Mohr and Gregor Snelting, "Checking Applications using Security APIs with JOANA", in 8th International Workshop on Analysis of Security APIs
15. Ferrante, Jeanne and Ottenstein, Karl J. and Warren, Joe D., "The Program Dependence Graph and Its Use in Optimization", in ACM Trans. Program. Lang. Syst., July 1987, https://doi.org/10.1145/24039.24041
16. Jurgen Graf and Martin Hecker and Martin Mohr, "Using JOANA for Information Flow Control in Java Programs - A Practical Guide", in Proceedings of the 6th Working Conference on Programming Languages (ATPS'13)
17. Leavens, Gary T and Baker, Albert L and Ruby, Clyde, "JML: a Java modeling language" in Formal Underpinnings of Java Workshop (at OOPSLA98)
18. Potter, Ben and Till, David and Sinclair, Jane, "An Introduction to Formal Specification and Z"
19. Lano, Kevin, "The B language and method: a guide to practical formal development"
20. A. Arnold, "Finite transition systems. Semantics of communicating systems"
21. Sabelfeld, Andrei and Myers, Andrew C, "Language-based information-flow security", in IEEE Journal on selected areas in communications
22. Mateescu, Radu and Thivolle, Damien, "A Model Checking Language for Concurrent Value-Passing Systems" in "Proceedings of the 15th International Symposium on Formal Methods" , https://doi.org/10.1007/978-3-540-68237-0_12
23. Garavel, Hubert and Lang, Frederic and Mateescu, Radu and Serwe, Wendelin "CADP 2010: A Toolbox for the Construction and Analysis of Distributed Processes" in "Tools and Algorithms for the Construction and Analysis of Systems: 17th International Conference, TACAS 2011", https://doi.org10.1007/978-3-642-19835-9_33
24. Flanagan, Cormac and Leino, K. Rustan M. and Lillibridge, Mark and Nelson, Greg and Saxe, James B. and Stata, Raymie, "PLDI 2002: Extended Static Checking for Java", http://doi.acm.org/10.1145/2502508.2502520
25. Chess, Brian V "Improving computer security using extended static checking"
26. De Francesco, Nicolette and Santone, Antonella and Tesei, Luca "Abstract Interpretation and Model Checking for Checking Secure Information Flow in Concurrent Systems"
27. Serge Autexier and Dieter Hutter and Bruno Langenstein and Heiko Mantel and Georg Rock and Axel Schairer and Werner Stephan and Roland Vogt and Andreas Wolpers, "VSE formal methods meet industrial needs", https://doi.org/10.1007/s100099900022
28. Akeel, Fatimah and Salehi Fathabadi, Asieh and Paci, Federica and Gravell, Andrew and Wills, Gary, "Formal Modelling of Data Integration Systems Security Policies", https://doi.org/10.1007/s41019-016-0016-y
29. John Wilander and et al, "Pattern Matching Security Properties of Code using Dependence Graphs"
30. Andrew, Johnson and Lucas, Waye and Scott, Moore, "Exploring and Enforcing Security Guarantees via Program Dependence Graphs", https://doi.org/10.1145/2737924.2737957

"Formal specification of security guidelines for program certification"

"Abstract Interpretation and Model Checking for Checking Secure Information Flow in Concurrent Systems" in Proceedings of Fundamenta Informaticae, 2003

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 49-65 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Improving System Reliability Assessment of Safety-Critical Systems using Machine Learning Optimization Techniques

Ibrahim Alagöz[*1], Thomas Hoiss[2], Reinhard German[1]

[1]*Department of Computer Science 7, FAU Erlangen-Nuremberg, 91058, Germany*

[2]*Automotive Safety Technologies GmbH, 85080, Gaimersheim, Germany*

ARTICLEINFO

Article history:
Received: 14 November, 2017
Accepted: 10 January, 2018
Online: 30 January, 2018

Keywords:
Safety-Critical System
Black Box Regression Testing
Linear Classifier
Selection
Prioritization

ABSTRACT

Quality assurance of modern-day safety-critical systems is continually facing new challenges with the increase in both the level of functionality they provide and their degree of interaction with their environment. We propose a novel selection method for black-box regression testing on the basis of machine learning techniques for increasing testing efficiency. Risk-aware selection decisions are performed on the basis of reliability estimations calculated during an online training session. In this way, significant reductions in testing time can be achieved in industrial projects without uncontrolled reduction in the quality of the regression test for assessing the actual system reliability.

1 Introduction

Reliability assessment of safety-critical systems is becoming an almost insurmountable challenge. In the near future, the engineering of new applications for vehicles such as driving assistance functions or even autonomous driving systems will inevitably incur significantly increased engineering sophistication and longer test cycles. Thus, in the automotive domain, functional safety continues to be ensured on the basis of the international ISO 26262 standard. As both the levels of functionality such systems provide and their degree of interaction with their environment increases, an adequate increase in system safety assessment capabilities is required.

This paper is an extension of work originally presented at the 10th IEEE International Conference on Software Testing, Verification and Validation (ICST 2017) [1] and describes a methodology for efficiently assessing system safety. The focus of the paper is on regression testing of safety-critical systems consisting of black-box components. This scenario is common for automotive electronic systems, where testing time is expensive and should be reduced without an uncontrolled reduction in reliability.

The work reported here correspondingly seeks to increase testing efficiency by reducing the number of selected test cases in a regression test cycle. When a selection decision is made, the following two types of errors are possible:

- a test case is selected but would pass (type-I-error, false-positive case) and
- a test case is not selected but would fail (type-II-error, false-negative case).

Accordingly, we model a classifier $\hat{H}$ for solving the following optimization problem.

$$\begin{aligned} &\min\ p_{FP} = P(\hat{H} = H_1|H_0) \\ &\text{subject to } p_{FN} = P(\hat{H} = H_0|H_1) \leq p_{FN,MAX} \end{aligned} \tag{1}$$

A good standard of test efficiency calls for the avoidance of false-positives. This requires minimization of the probability of mistakenly assuming the rival hypothesis (H_1 : test case fails) even though the null hypothesis (H_0 : test case passes) is correct. Conversely, false-negatives mean that system failures remain undetected; the occurrence of this type of error must therefore be avoided with very stringent requirements. Thus, a predefined limit $p_{FN,MAX}$ for the probability of a false-negative is defined.

[1] proposed a concept for the selection of test cases based on a stochastic model. However, this paper proposes a holistic optimization framework for the safety assessment of safety-critical systems based on machine learning optimization techniques. We

[*]Corresponding Author: Ibrahim Alagöz, Hornstr. 1 85051 Ingolstadt, ibrahim.alagoez@gmail.com

https://dx.doi.org/10.25046/aj030107

suggest an incrementally and actively learning linear classifier whose parameters are estimated on the basis of Bayesian inference rules. As a result, our novel approach for modeling a linear classifier outperforms other machine learning approaches in terms of sensitivity.

Furthermore, this paper deals with the following fundamentally important research question: The machine learning approach is trained with data (test evaluations) obtained during a concurrently running regression test. How much training data is enough? When does regression test selection actually start?

We extend the proposed selection method [1] by introducing suitable test case features that are used in the machine learning approach for increasing performance (see [2]). Therefore, each feature introduced increases the complexity of the optimization problem (cf. Eq. 1) as a new dimension for optimization is introduced. Thus [3] and [4] suggest that high dimensional optimization problems can be solved in reasonable timeframes by using evolutionary algorithms instead of a (grid)search-based approach as given in [1]. Accordingly, we propose an evolutionary optimization approach for increasing testing efficiency.

Further extensions, such as the introduction of a prioritization strategy for test cases in order to select higher-priority test cases, will also be presented within this paper. In our novel approach, a linear classifier is trained in an online session; the ordering of the training data on the basis of a prioritization strategy therefore has the potential to improve our classifiers' performance.

We also provide an industrial case study to show the advantages of the suggested selection method. The study uses data from several regression test cycles of an ECU of a German car manufacturer, showing how test effort can be reduced significantly whereas the rates of both false-negatives and false-positives can be kept at very low values. In this example, we can quadruple test efficiency by keeping the false-negative probability at 1%.

We first discuss related work in Sec. 2 and explain basic definitions in Sec. 3. Accordingly, we motivate our research topic in Sec. 4 by giving some background information on regression tests and referring to the challenges. In Sec. 5, we give a brief overview of known machine learning methods' performance in solving safety-critical binary classification tasks. The concept of our novel machine learning approach is presented in Sec. 6. Sec. 7 discusses optimization strategies, and Sec. 8 focuses on the importance of the learning phase for the success of our approach. An industrial case study with real data is then given in Sec. 9. Finally, Sec. 10 presents the paper's conclusions.

2 Related Work

The automotive industry is currently engaged in a laborious quality assessment process around new engineered driver assistance and active and passive safety functions, while functional safety is ensured according to the international ISO 26262 standard [5]. Reliability assessment of systems is therefore, possible through both model-checking and testing.

Model-checking is used for verifying conditions on system properties. Thus [6] states that system requirements can also be validated by model-checking techniques. The idea is to check the degree to which system properties are met and to deduce logical conclusions on the basis of the satisfaction of system requirements. Model-checking has therefore gained wide acceptance in the field of hardware and protocol verification communities [7]. Motivated by the fact that numerical model-checking approaches cannot be directly applied to black-box components as a usable formal model is not available, we focus on model-checking driven black-box testing [6] and statistical model-checking techniques [8]. However, there exist some approaches for interactively learning finite state systems of black-box components (see [9] and [10]), which are proposed as *black box checking* in [11]. Learning a model is an expensive task, as the interactively learned model has to be adapted due to inaccuracy reasons. Nevertheless, some assumptions about the system to be checked, such as the number of internal states, are necessary; furthermore, conformance testing for ensuring the accuracy of the learned model has to be iteratively performed [9].

Therefore, [8] outlines the advantages of statistical model-checking as being simple, efficient and uniformly applicable to white- and even to black-box systems. [6] motivates on-the-fly generation of test cases for checking system properties; here, a test case is generated for simulating a system for a finite number of executions. All these executions are used as individual attempts to discharge a statistical hypothesis test and finally for checking the satisfaction of a dedicated system property.

Model-checking driven testing, or even simply testing a system in order to validate its requirements, is an expensive task, especially where safety-critical systems are concerned. However, the focus is on regression testing, which means that the entire system under test has already been tested once but has to be tested again due to system modifications that have been carried out. The purpose of regression testing is to provide confidence that unchanged parts within the system are not affected by these modifications [12]. White-box selection techniques have been comprehensively researched [13, 14]. However, we are here considering black-box components, and hence selecting test cases that only check modified system blocks gets difficult.

Since the implementation of black-box systems and moreover, the information on performed system modifications is not available [12], reasonably conducting a regression test becomes impossible.

Accordingly, regression testing of safety-critical black-box systems ends up in simply executing all existing test cases; this is a *retest-all* approach [12].

This is also motivated by the fact that in the au-

tomotive industry, up to 80% of system failures [1] that are detected during a regression test have not occurred previously. The reason behind this fact is that often many unintended bugs are introduced during a bug-fixing process. So between two system releases many new unknown errors are often introduced.

For reducing the overall test effort, we apply a test case selection method [1] based on hypothesis tests. Those test cases that are assumed to fail on their executions are accordingly selected. However, type errors while performing hypothesis tests are possible, as, for instance, in statistical model-checking.

We extend the proposed selection method into a holistic machine learning-driven optimization framework that utilizes suitable test case features for increasing testing efficiency (see [2]). Machine learning methods are often trained in so-called *batch* modes. Nevertheless, many applications in the field of autonomous robotics or driving are trained on the basis of continuously arriving training data [15]. Thus, incremental learning facilitates learning from streaming data and hence is exposed to continuous model adaptation [15]. Especially handling non-stationary data assumes key importance in applications like voice and face recognition due to dynamically evolving patterns. Accordingly, many adaptive clustering models have been proposed, including incremental K-means and evolutionary spectral clustering techniques [16].

Furthermore, labeling input data is often awkward and expensive [17] and hence accurately training models can be difficult. Therefore, semi-supervised learning techniques are developed for learning from both labeled and unlabeled data [17]. Motivated by these techniques, we propose a similar approach for effectively learning from labeled data. Hence, we cluster binary labeled data in more than two clusters for improving a classifier's learning capability due to the optimization of an objective function. Our optimization framework thus utilizes evolutionary optimization algorithms for handling the optimization complexity. Minimization of labeling cost on the basis of active learning strategies [18, 19] will also be dealt with in this paper.

3 Basic Definitions

We define the test suite $T = \{t_i \mid 1 \leq i \leq M\}$ consisting of a total of M test cases. $T_{Exec} \in T$ and $T_{\overline{Exec}} \in T$ are subsets of T that contain test cases that are executed and deselected in a current regression test respectively. Based on the test case executions ($\forall t_i \in T_{Exec}$), a system's reliability is actually learned, and thus the machine learning algorithm is trained.

The focus in supervised learning is on understanding the relationship between feature and data (here test case evaluation) [4]. Therefore, a test case needs to code a feature vector so that the indication of the coded features for a system failure can be learned in a supervised fashion. Such an indication is not just a highly probable forecast of an expected system failure, it is rather a particular risk-associated recognition.

First of all, a feature can be any individual measurable property of a test case. The data type of a feature is mostly numeric, but strings are also possible. However, such features need to be informative, discriminative and independent of one another if they are to be relevant and non-redundant The definition of suitable features increases the classifier performance [20]. In our application, a feature can be varied, such as a

- subjective ranking of a test case based on expert knowledge. Such rankings can hint at the error susceptibility of verified parts of the system;
- verified function's safety integrity level, known as the *ASIL* in automotive applications [5];
- name of a function whose reliability is assured;
- reference to any hardware component of a circuit board that is being tested in a hardware-in-the-loop (HiL) test environment;
- number of totally involved electronic control units during the testing of a networked functionality; Such a number can hint at the complexity of the networked functionality and hence at its error susceptibility.

We define the entire set of features $\Phi = \{\phi_f \mid 1 \leq f \leq F\}$ of test cases that might be relevant for understanding the behavior of test cases. Thus, features may be e.g. $\phi_1 = \{'QM','A','B','C','D'\}$ ($ASIL$) or $\phi_2 = \{f_1, f_2, f_3\}$ (function name). Hence, a test case can verify a function f_3 that has an *ASIL* A.

The following passages discuss the selection of suitable features, which is an important strategy for improving a classifier's performance.

- Sometimes less is more - If the defined set Φ is too large it can cause huge training effort, high dimensionality of the optimization problem and overfitting. Thus we define a selection mask $\boldsymbol{b}_s = \begin{bmatrix}1 & 0 & 0 \cdots 1\end{bmatrix}$ of length F for selecting relevant features Φ_s. If the $f-th$ matrix entry of $\boldsymbol{b}_s$ is greater than or equal to 1, then the corresponding feature $\phi_f \in \Phi$ is selected and added to Φ_s, otherwise not.
- The set of main features $\Phi_m \subseteq \Phi_s$ is coded as follows: If the $f-th$ matrix entry of $\boldsymbol{b}_s$ is equal to 2, then the corresponding feature $\phi_f \in \Phi_s$ is at the same time a main feature $\phi_f \in \Phi_m$, otherwise not. The main features are used to establish the overall training data set: The training data is adapted to each test case, and thus it is $T_{t_i} = \{t_j \mid t_j \in T_{Exec} \wedge t_i \overset{\Phi_m}{\equiv} t_j\}$. Hence, we define that two test cases t_i and t_j are equivalent $t_i \overset{\Phi}{\equiv} t_j$ if their features $\forall \phi_f \in \Phi$ have identical values.

Additionally, a cross-product transformation of features $\forall \phi_f \in \Phi_s \setminus \Phi_m$ is performed. Thus we define $\Psi = \phi_{f_1} \times \phi_{f_2} \times ... \times \phi_{f_h} \times ... \times \phi_{f_H}$ with $\phi_{f_h} \in$

$\Phi_s \setminus \Phi_m, 1 \leq h \leq H = |\Phi_s \setminus \Phi_m|$ consisting of features ψ_l that represent individual combinational settings for features $\phi_f \in \Phi_s \setminus \Phi_m$. In simple terms, the cross-product of our sample features is $\phi_1 \times \phi_2 = \{('QM', f_1), ('QM', f_2), ('QM', f_3), ('A', f_1), ...\}$ Finally, a Boolean function $check : T \times \Psi \rightarrow \mathbb{B}$ with $check(t_i, \psi_l) = \begin{cases} 0, & \text{if } t_i\text{'s features are given by } \psi_l \\ 1, & otherwise \end{cases}$ is defined.

In addition, the function $state$: $T \times R \rightarrow S$ is defined; it returns the state of a dedicated test case in a concrete regression test. The state has to be either 'Pass' or 'Fail', except for cases where the test case has not been executed so that its state is undefined. Therefore, $S = \{'Pass', 'Fail', 'Undefined'\}$ defines the set of possible states. Furthermore, the set $R = \{r_k \mid 0 \leq k \leq K\}$ includes r_0 which is the current regression test and older regression tests starting from the last regression r_1 to the first considered regression r_K. Lastly, we define the tuple $history(t_i) = \{state(t_i, r_1), ..., state(t_i, r_K)\}$ containing t_i's previous test results.

4 Motivation

In practice, finding suitable features is a difficult task. Since we focus here on black-box systems, system-internal information is not available that might be useful for understanding the system behavior. As a reason, we can only define the above listed features, which might be too high-level for classifying system failures. To illustrate this fact, Fig. 1 a) shows a typical situation: The behavior of test cases in relation to arbitrarily defined features ϕ_1 and ϕ_2 is given. Passed and failed test cases are presented by green squares and red diamonds respectively, and white circles stand for test cases yet to be executed.

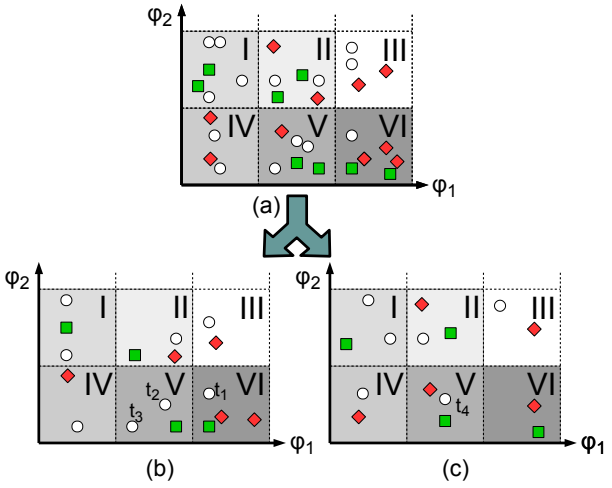


Figure 1: An artificial regression test with test cases.

We can see that the green squares and the red diamonds are widely scattered. Thus, defining a hyperplane in order to set two acceptance regions for passing and failing test cases is no easy matter. In order to solve this complex task, we develop a novel approach that is basically motivated by the following thought experiment: All test cases that are represented in Fig. 1 a) are now either assigned to 1 b) or 1 c) according to a certain mapping. The individual mappings of test cases will be discussed later, in Sec. 6. In the next step, a cross-product transformation is performed in order to group test cases into sub-regions (we refer these later as sub-clusters). In our example, we create six sub-regions. Table 1 lists the empirical failure probabilities of each sub-region.

In order to keep our thought experiment very simple, we will neglect statistical computations for now and focus only on the main idea of our novel approach. The introduction of Bayesian networks and hence the derivation of weights for linear classifiers will be discussed later, in Sec. 6. We assume for now that the calculated failure probabilities of test cases in Fig. 1 b) and Fig. 1 c) are correlated. So our example remains very simple, we also require that the failure probabilities of the corresponding sub-regions are equal. This assumption reduces the complexity of the following classification task enormously. We will classify the following test cases t_1, t_2, t_3 and t_4 in accordance with whether a selection is necessary or not.

Table 1: Failure probability of each sub-region.

$P(H_1)$	I	II	III	IV	V	VI
Fig. 1 a)	0	0.5	1	1	1/3	3/5
Fig. 1 b)	0	0.5	1	1	0	2/3
Fig. 1 c)	0	0.5	1	1	0.5	0.5

Only if t_1 passes will the failure probability of sub-region VI in Fig. 1 b) be equal to the failure probability of sub-region VI in Fig. 1 c). According to this fact, t_1 is assumed to pass, and hence it is deselected. Furthermore, t_2 will be selected as a fail of a test case inside sub-region V is expected. However, t_2 passes, and, based on the same consideration, t_3 is also selected and finally fails. Since now a failure probability of 1/3 is expected in sub-region V, t_4 is assumed to pass, and therefore it does not need to be selected. Table 2 summarizes all decisions executed.

Table 2: Test case states and algorithm decisions.

Test Case	State	Decision	Type of Decision
t_1	Pass	Deselected	True-Negative
t_2	Pass	Selected	False-Positive
t_3	Fail	Selected	True-Positive
t_4	Pass	Deselected	True-Negative

So our novel approach for solving binary classification tasks is based on calculated empirical probabilities and empirically evaluated correlations among those probabilities. The behavior of test cases can be

precisely estimated on the basis of the calculated correlations. In practice, the failure probabilities of same sub-regions in Fig. 1 b) and Fig. 1 c) is often not exactly equal, but these failure probabilities are correlated. So the main task is to find *good* sub-regions for maximizing the empirically evaluated correlations and thus for precisely estimating the behavior of test cases. A more detailed explanation of our novel approach will follow in Sec. 6.

5 Performance of Known Machine Learning Methods

We have already indicated, by showing the example regression test in Sec. 4 (see Fig. 1), that according to the distribution of the input data, many machine learning methods cannot be reasonably applied for solving the constrained optimization problem (cf. Eq. 1). We will now demonstrate briefly that training linear classifiers in the classical sense by minimizing a loss function cannot perform well for solving safety-critical binary classification tasks. The situation is that only a small percentage of the data is actually labeled with one ('Fail'). Furthermore, failed and passed test cases are widely scattered in the feature space, which means detecting failing test cases becomes impossible. Additionally, the performance of deep neuronal networks is validated in the following.

The evaluation results (precision/recall) of these machine learning methods are given in Table 3. Each machine learning method is trained in the *batch* mode. The training data consists of all obtained test evaluations of a special regression test that will also be analyzed in our industrial case study in Sec. 9. For evaluating the machine learning methods, we used the training data first for training and later for testing (training data = test data). Even so, the sensitivity of both machine learning methods is zero, and thus we propose a novel approach for determining a linear classifier's parameters in Sec. 6.

Table 3: Performance of known machine learning methods in solving safety-critical binary classification tasks.

Machine Learning	Precision	Recall / Sensitivity
Linear Classifier (Trained by Minimizing a Loss-Function)	0	0
Deep Neuronal Network	0	0

6 Concept

The concept of our novel approach is shown in Fig. 2. We start with specifying a feature set Φ, and taking its subset Φ_s and finally constitute a cross-product feature transformation to obtain the set Ψ. Based on Ψ and by applying the *check*-function on T_{Exec}, test cases can be grouped. If we look back to the example where we grouped test cases in Fig. 1 a), then we will see that there is a relationship between test cases' features and their assignments to sub-regions.

Correspondingly, we introduce the definitions of clusters and sub-clusters of test cases. In the first step, test cases $\forall t_k \in T_{t_i}$ are assigned to clusters based on their history-tuples. A cluster is basically a partition of T_{t_i} and consists of test cases that have the same history-tuples. Accordingly, the number of distinct history-tuples N determines the total number of clusters. This is the step that has already been shown in Sec. 4, where test cases inside Fig. 1 a) were individually mapped into Figs. 1 b) and 1 c). In this way, already executed test cases depicted in Fig. 1 b) belong to one cluster and, analogously, those executed test cases that are depicted in Fig. 1 c) belong to another cluster.

In the next step, each cluster C_n is subdivided into L sub-clusters. A test case $t_k \in C_n$ is an element of the l-th sub-cluster $C_{n,l}$ if $check(t_k, \psi_l)$ is true. We originally introduced the terminology of sub-regions in Sec. 4. However, we focus in what follows on discrete valued features, which means that grouping test cases into sub-clusters is more appropriate. By introducing the function $eval : T_{Exec} \rightarrow \{0, 1\}$ that is defined as follows

$$eval(t_i) = \begin{cases} 0, & \text{for } state\{t_i, r_0\} = \text{'Pass'} \quad (2) \\ 1, & \text{for } state\{t_i, r_0\} = \text{'Fail'} \quad (3) \end{cases}$$

the calculation of failure probabilities can be given in Eq. 4.

$$p_{n,l} = \frac{1}{|C_{n,l}|} \sum_{\forall t_i \in C_{n,l}} eval(t_i) \quad (4)$$

The given selection decisions in our example in Sec. 4 (cf. Fig. 1) were taken based on calculated failure probabilities. Additionally, the correlations between the failure probabilities were considered. Accordingly, we need a stochastic model for estimating the classifier's sensitivity and specificity. We propose a univariate and also a multivariate stochastic model. The short-comings of the univariate stochastic model for solving the optimization problem (cf. Eq. 1) will be discussed later to motivate the introduction of a multivariate stochastic model. First of all, the next step introduces a multidimensional Gaussian distribution that constitutes a distribution for the failure probabilities of test cases. Based on this distribution, two distinct Bayesian Belief networks for both stochastic models will be introduced.

In the following, we interpret $p_{n,l}, 1 \leq l \leq L$ as realizations of a random variable X_n. X_n is Gaussian distributed based on the following assumption:

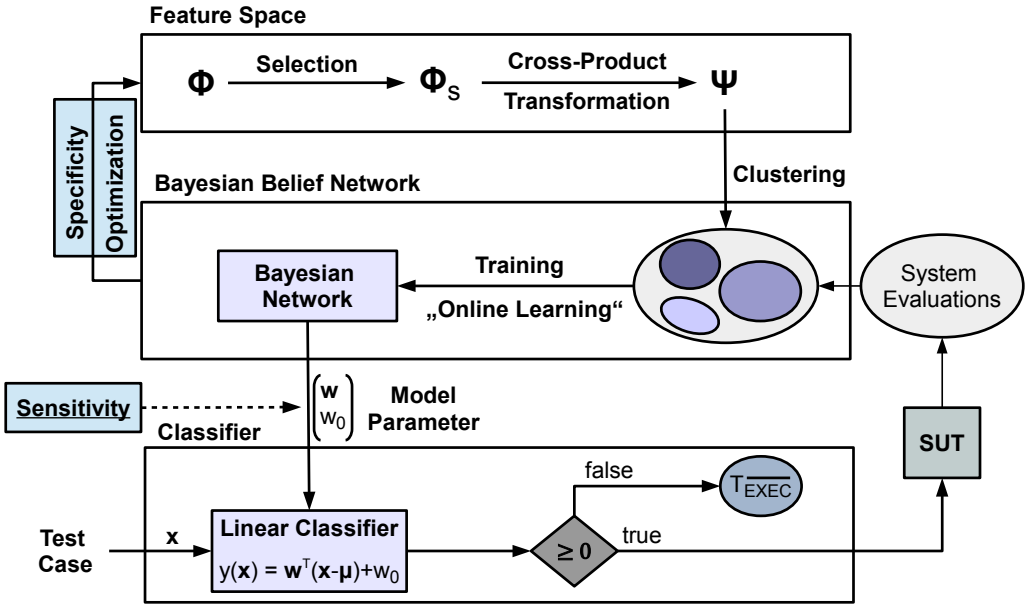


Figure 2: Determining the weights of a linear classifier for maximizing its specificity under the constraint of a specific sensitivity.

Since each test case evaluation is a binary experiment with two possible outcomes ('0' or '1'), it can be regarded as a realization of a binary random variable. As the sum of independent random variables results into a Gaussian random variable according to the central-limit theorem [21], considering test case evaluations as independent random experiments justifies X_n's assumed distribution. However, test cases are executed on the same system and there may be some dependencies between test case evaluations that cannot be directly validated by such means as performing code inspections. As a result, we assume a mix of dependent and independent test case evaluations, and hence the Gaussian assumption is still valid. The moments of X_n are $E[X_n] = \mu_n = \frac{1}{L}\sum_{l=1}^{L} p_{n,l}$ and $E[(X_n - E[X_n])^2] = \sigma_n^2 = \frac{1}{L-1}\sum_{l=1}^{L}(p_{n,l} - E[X_n])^2$. As we introduced in total N Gaussian random variables, the moments of the multidimensional Gaussian distribution are $\boldsymbol{\mu} = E[\boldsymbol{X}] = [E[X_1], E[X_2], ..., E[X_N]]^T$ and $\Sigma = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]$.

Since the constraint of the optimization problem (cf. Eq. 1) has to be fulfilled, an accurate sensitivity estimation has to be iteratively performed.

6.1 Sensitivity Estimation

The formula for calculating the classifier's false-negative selection probability is given in Eq. 5.

$$p_{FN} = \frac{N_{FN}}{N_{FN} + N_{TP}} \leq p_{FN,MAX} \tag{5}$$

However, $\hat{p}_{FN} = \frac{\hat{N}_{FN}}{\hat{N}_{FN}+N_{TP}}$ has to be estimated, since the number of mistakenly deselected failing test cases N_{FN} is unknown, and thus it is estimated by $\hat{N}_{FN}$. The number of already detected failing test cases is given by N_{TP}. Before a decision can be taken on whether a test case t_i can be deselected, the currently allowed risk of taking a wrong decision has to be estimated in advance. The estimation of $\hat{N}_{FN}$ has to be adjusted by the term $xP(\hat{H} = H_0|H_1)$, where x is the failure probability of t_i and $P(\hat{H} = H_0|H_1)$ is the estimated false-negative probability if t_i is deselected. Accordingly, the recursive formulation $\hat{N}_{FN,new} = \hat{N}_{FN,old} + xP(\hat{H} = H_0|H_1)$ is continuously updated whenever an arbitrary test case is deselected. As $\hat{p}_{FN} \leq p_{FN,MAX} \Leftrightarrow \frac{\hat{N}_{FN,new}}{\hat{N}_{FN,new}+N_{TP}} \leq p_{FN,MAX}$ is required, the maximum allowed false-negative probability for deselecting the next test case t_i is given in Eq. 6.

$$\begin{aligned} P(\hat{H} = H_0|H_1) &\leq \frac{\frac{N_{TP} \cdot p_{FN,MAX}}{1-p_{FN,MAX}} - \hat{N}_{FN,old}}{x} \\ &=: p_{FN,Limit} =: \frac{p_{FN,Bound}}{x} \end{aligned} \tag{6}$$

6.2 Univariate Stochastic Model

We model the Bayesian Network that consists of the random variables X, H and $\hat{H}$, in Fig. 3. In the univariate stochastic model, the focus is on modeling of only one failure probability distribution. Thus the random variable X stands for the previously defined X_1 and its realization x is given by $p_{1,l}$ where l is the index of that sub-cluster $C_{1,l}$ that fulfills $t_i \in C_{1,l}$.

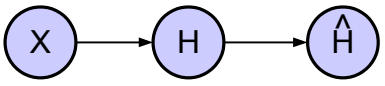


Figure 3: Bayesian Network consisting of the random variables X, H and $\hat{H}$.

H and $\hat{H}$ are binary random variables for modeling

test case states and classifier decisions. As the state of a test case t_i is a-priori unknown, it needs to be modeled by a corresponding random variable. According to the realization of X, a pass or a fail of the corresponding test case t_i, whose failure probability distribution is modeled by X, is expected. Finally, $\hat{H}$ takes a decision for t_i based on its failure probability x:

$$\hat{H}(x) = \begin{cases} H_0, & \text{if } x \in \mathcal{X}_0 = [0; p_{TH}[\quad (7) \\ H_1, & \text{if } x \in \mathcal{X}_1 = [p_{TH}; 1] \quad (8) \end{cases}$$

According to $\hat{H}$'s selection rule a very simple *hyperplane* $y(x) = x - p_{TH}$ is derived where in the case of $y(x) \geq 0$ a selection decision is taken. A particularly important factor is the definition of the threshold probability p_{TH}, as its setting determines the classifier's sensitivity and specificity. The common ways of estimating false-negative and false-positive probabilities are given in equations 9 and 10, respectively.

$$\hat{p}_{FN} = \int_{\mathcal{X}_0} p(X|H_1)dx \quad (9)$$

However, we could only estimate the probability density function (pdf) $p(X)$ in contrast to the conditional probability density functions $p(X|H_0)$ and $p(X|H_1)$. The reason for this is that pdf estimations are based on mean calculations of test case evaluations. Hence passing and failing test cases are both considered in calculating average failure probabilities. Thus, $p(X)$ is a distribution over failure probabilities of passing as well as failing test cases. Accordingly, $p(X|H_0)$ and $p(X|H_1)$ cannot be estimated and in conclusion, $\hat{p}_{FN}$ and $\hat{p}_{FP}$ cannot be estimated as in Eq. 9 and 10.

$$\hat{p}_{FP} = \int_{\mathcal{X}_1} p(X|H_0)dx \quad (10)$$

Fig. 4 shows the important probability distribution functions that are used for estimating $\hat{p}_{FN}$ and $\hat{p}_{FP}$.

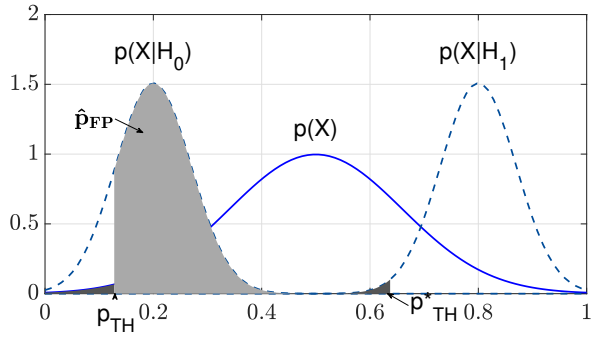


Figure 4: Considered probability distribution functions in the univariate stochastic model.

The threshold probability p_{TH} is calculated based on the estimation of $\hat{p}_{FN}$ as the following relation in Eq. 11 holds.

$$\hat{p}_{FN} = P(X < p^*_{TH}|H_1) \quad (11)$$

As Eq. 11 cannot be directly estimated, the following relation in Eq. 12 is used for estimating $\hat{p}_{FN}$ and finally for p_{TH}.

$$\hat{p}_{FN} = P(X < p_{TH}|H_1) \leq P(X < p_{TH}) \leq p_{FN,Limit} \quad (12)$$

$\hat{p}_{FN}$ is estimated in Eq. 12 according to the assumption that the quantiles of $p(X|H_1)$ are larger than the quantiles of $p(X)$. By solving Eq. 12 the threshold probability is computed as given in Eq. 13

$$p_{TH} = \text{erfinv}(2p_{FN,Limit} - 1)\sigma\sqrt{2} + \mu \quad (13)$$

with $\mu = E[X]$ and $\sigma = \sqrt{VAR(X)}$. As a result, the classifier's sensitivity is larger than $1 - p_{FN,Limit}$, since its false-negative selection probability is smaller than $p_{FN,Limit}$. Finally, the decision regions of the linear classifier are defined (cf. Eq. 7 and 8) by determining p_{TH}.

Furthermore, the minimization of the classifier's specificity is required by the definition of the constraint optimization problem (cf. Eq. 1). Accordingly, the classifier's false-positive selection probability is estimated as given in Eq. 14 and shown in Fig. 4.

$$\hat{p}_{FP} = P(X \geq p_{TH}|H_0) \quad (14)$$

However, $p(X|H_0)$ is not given and thus $\hat{p}_{FP}$ cannot reasonably be estimated. Furthermore, $\hat{p}_{FP}$ cannot be reasonably minimized as an optimization parameter is not defined; consequently, we need a so-called multivariate stochastic model for performing this. In the first instance, the idea of minimizing $\hat{p}_{FP}$ and hence gaining testing efficiency by regarding several distribution functions is explained.

6.3 Preliminaries

Let us assume that two dependent Gaussian random variables X and X' are given. The focus is again on estimating $\hat{p}_{FN}$ and $\hat{p}_{FP}$. Fig. 5 shows the probability distribution functions $p(X)$, $p(X|H_0)$ and $p(X|H_1)$ as in Fig. 4. Additionally, the a-posteriori failure probability distribution function $p(X|X')$ is shown.

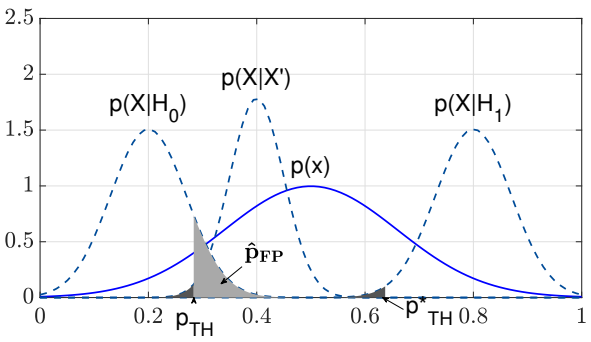


Figure 5: Qualitatively minimizing $\hat{p}_{FP}$ by introducing $p(X|X')$ and hence by using conditional informations.

The main idea is to use several observations of distinct dependent random variables to achieve a consid-

erably more representative a-posteriori failure probability distribution function that is relatively narrow within a certain range. So $p(X|X')$ is considered as the more representative distribution for the failure probabilities and hence $\hat{p}_{FN}$ and p_{TH} are estimated by using this distribution function. Comparing Figs. 4 and 5, it can easily be seen that $\hat{p}_{FP}$ is basically minimized, since the risk of a false-negative selection probability is computed based on $p(X|X')$, which allows a more representative risk estimation.

All in all, by regarding a set of dependent Gaussian random variables and by using the information about their observations, a more representative a-posteriori failure probability distribution function is achieved, which allows a more precise risk estimation. Accordingly, the probability of false-positive selection can be minimized. As a result, a multivariate stochastic model is created to exploit the dependency information between random variables for finally achieving testing efficiency.

6.4 Multivariate Stochastic Model

By using the dependency between the random variables $X_n, 1 \leq n \leq N$, a considerably more accurate estimation of $\hat{p}_{FN}$ is achieved and hence $\hat{p}_{FP}$ is minimized. Fig. 6 shows the modeled Bayesian network consisting of the random variables $X_n, 1 \leq n \leq N$, H and $\hat{H}$.

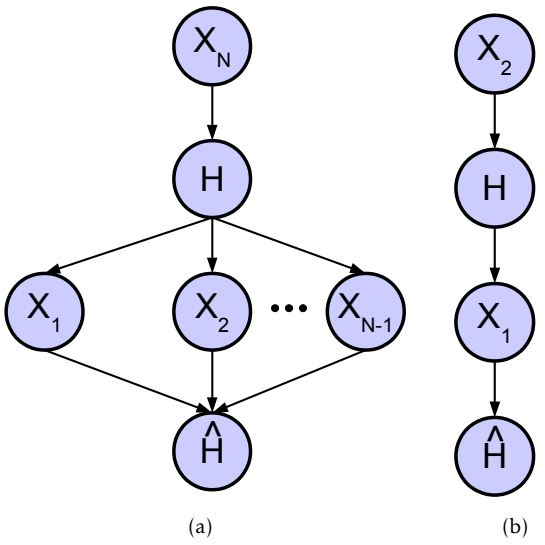


Figure 6: Bayesian Network consisting of an (a) indefinite and (b) definite number of random variables $X_n, 1 \leq n \leq N$.

The focus is on taking a selection decision for an arbitrary test case t_i. X_N now models the failure probability distribution of t_i. In our previous example, as shown in Fig. 1 c), the failure probability distribution of test case t_4 was calculated based on the empirically evaluated failure probabilities of test cases inside Fig. 1 c). Thus t_4 was an element of C_2 ($N = 2$), and its failure probability was modeled by X_2. Analogously, X_1 was defined by the empirically evaluated failure probabilities of test cases inside Fig. 1 b). In the interest of simplification, we always assume that the currently focused test case t_i is an element of cluster C_N and thus X_N models its failure probability distribution.

Furthermore, we can calculate the dependency among $X_n, 1 \leq n \leq N$. However, the Bayesian network in Fig. 6 models the statistical dependency between H and further random variables $X_n, 1 \leq n \leq N-1$. These dependencies cannot be calculated, but have to be modeled for estimating $\hat{p}_{FN}$ and $\hat{p}_{FP}$.

First of all, we model the classifier $\hat{H}(\cdot)$ as follows

$$\hat{H}(x_{ML}) = \begin{cases} H_0, & \text{if } x_{ML} \in \mathcal{X}_0 = [0; p_{TH}[\quad (15) \\ H_1, & \text{if } x_{ML} \in \mathcal{X}_1 = [p_{TH}; 1] \quad (16) \end{cases}$$

where $x_{ML} = \underset{x_N}{\operatorname{argmax}}\Big[\ln[\mathcal{L}(x_N|x_1,...,x_{N-1})]\Big]$ (consult [1]) is the maximum likelihood estimation. Accordingly, the likelihood estimation is a weighted sum as given in Eq. 17

$$x_{ML} =: \sum_{n=1}^{N-1} w_n(x_n - \mu_n) + \mu_N \tag{17}$$

with weights $w_n, 1 \leq n \leq N-1$ as given in Eq. 18.

$$w_n = -\frac{(\Sigma^{-1})_{n,N}}{(\Sigma^{-1})_{N,N}} \tag{18}$$

Further, p_{TH} has to be calculated based on a precise estimation of $\hat{p}_{FN}$. Thus we derive a calculation formula for $\hat{p}_{FN}$ for the case $N = 2$, but we will also provide a general calculation formula of $\hat{p}_{FN}$ for an arbitrary number N of random variables.

6.4.1 Derivation of probability distribution functions

In the following, some probability distributions are driven that are used for estimating $\hat{p}_{FN}$. First of all, the joint pdf $p(\hat{H}X_1HX_2)$

$$p(\hat{H}X_1HX_2) = p(\hat{H}|X_1)p(X_1|H)p(H|X_2)p(X_2) \tag{19}$$

and the conditional pdf $p(\hat{H}X_1H|X_2)$ are given in Eq. 19 and 20, respectively.

$$\begin{aligned} p(\hat{H}X_1H|X_2) &= \frac{p(\hat{H}X_1HX_2)}{p(X_2)} \\ &= \underbrace{p(\hat{H}|X_1)p(X_1|H)}_{p(\hat{H}X_1|H)}p(H|X_2) \end{aligned} \tag{20}$$

In the next step Eq. 21 is obtained by setting the equation $p(\hat{H}|X_1)p(X_1|H) = p(\hat{H}X_1|H)$ into Eq. 20.

$$p(\hat{H}X_1|H) = \frac{p(\hat{H}X_1H|X_2)}{p(H|X_2)} \tag{21}$$

Furthermore, the most important relation $p(\hat{H}|H) \leq \frac{p(\hat{H}|X_2)}{p(H|X_2)}$ is driven in Eq. 22.

$$p(\hat{H}X_1|H) \leq p(\hat{H}|H) \leq \frac{p(\hat{H}H|X_2)}{p(H|X_2)} \leq \frac{p(\hat{H}|X_2)}{p(H|X_2)} \tag{22}$$

Thus, the probability calculation $P(\hat{H} = H_0|H_1)$ can be estimated by using the relation in Eq. 22 as given in Eq. 23.

$$\hat{p}_{FN} = P(\hat{H} = H_0|H_1) \leq \frac{P(\hat{H} = H_0|X_2 = x_2)}{P(H_1|X_2 = x_2)} \tag{23}$$

Since $\hat{p}_{FN}$ cannot be directly estimated, as the conditional pdf $p(\hat{H}|H)$ is not given for performing the probability calculation $P(\hat{H} = H_0|H_1)$, the relation in Eq. 23 is used for estimating an upper bound for $\hat{p}_{FN}$. However, the linear classifier's actual false-negative deselection probability would be smaller than the calculated upper bound.

Since the constraint in Eq. 24 has to be fulfilled,

$$P(\hat{H} = H_0|H_1) \leq p_{FN,Limit} \tag{24}$$

we solve the inequality in Eq. 25.

$$\frac{P(\hat{H} = H_0|X_2 = x_2)}{P(H_1|X_2 = x_2)} \leq p_{FN,Limit} \tag{25}$$

As $x_2 = P(H_1|X_2 = x_2)$ holds, the following inequality is finally solved.

$$P(\hat{H} = H_0|X_2 = x_2) \leq p_{FN,Bound} \tag{26}$$

Eq. 26 is driven for the case $N = 2$ but in the general case, where the number of random variables $X_n, 1 \leq n \leq N$ is given by an arbitrary N, the following inequality has to be solved.

$$P(\hat{H} = H_0|X_N = x_N) \leq p_{FN,Bound} \tag{27}$$

By solving Eq. 27, the threshold probability

$$p_{TH} = -w_N(x_N - \mu_N) - w_0 + \mu_N \tag{28}$$

is obtained with weights

$$w_N = -\frac{e^{2I} - 1}{e^{2I}} \tag{29}$$

and

$$w_0 = -\frac{\sqrt{2}\sigma_N\sqrt{e^{2I} - 1}\,\mathrm{erfinv}(2p_{FN,Bound} - 1)}{e^{2I}} \tag{30}$$

Thus, the differential mutual information is defined in Eq. 31.

$$\mathcal{I} := \mathcal{I}(X_1, ..., X_{N-1}; X_N) \tag{31}$$

6.4.2 Conditional Independence

We have already motivated and introduced the following dependent random variables $X_n, 1 \leq n \leq N$. We have explained the fact that test case failure probabilities are correlated, since test cases are executed on the same system, and thus they show a dependent behavior.

However, the random variables $X_n, 1 \leq n \leq N$ are conditionally independent. This means that the information about a test case evaluation dominates such that a test case's originally calculated failure probability becomes irrelevant after observation of its state. Accordingly, the dependency among failure probabilities vanishes after observation of test case evaluations. This means that a fail of a test case t_m is actually expected based on the information about the evaluation of another test case t_n and no longer on t_n's originally calculated failure probability. Thus the remaining random variables $X_n, 1 \leq n \leq N-1$ become independent of the random variable X_N after observation of H's realization (cf. Fig. 6).

6.4.3 Specificity Estimation

The specificity is given by the term $1 - \hat{p}_{FP}$. As extensive mathematical derivations are needed for obtaining a calculation formula of $\hat{p}_{FP}$, these derivation steps are given in the appendix and in what follows here only the result is given.

Theorem 1 (False-Positive Probability Estimation)**.** *$\hat{p}_{FP}$ is estimated as given in Eq. 32*

$$\hat{p}_{FP} = P(Z \geq -w_0 + \boldsymbol{w}^T \Delta\boldsymbol{\mu}) = \frac{1}{2}\left[1 - erf\left(\frac{-w_0 + \boldsymbol{w}^T \Delta\boldsymbol{\mu}}{\sqrt{2}\sigma_Z}\right)\right] \tag{32}$$

with $\sigma_Z = [w_1 \cdots w_{N-1}]\Sigma_{1,1}[w_1 \cdots w_{N-1}]^T + w_N^2\sigma_N^2$ and $\Delta\boldsymbol{\mu} = \boldsymbol{\mu} - \boldsymbol{\mu}_{H_0}$. The conditional moment has the following definition $E[\boldsymbol{X}|H_0] = \boldsymbol{\mu}_{H_0}$

For the case $N = 2$, Eq. 32 can be simplified; after several calculation steps the following Eq. 54 results

$$\hat{p}_{FP} = \frac{1}{2}\left[1 - erf\left(\psi\right)\right] \tag{33}$$

with

$$\psi = \frac{\sqrt{e^{2\mathcal{I}} - 1}\,erfinv(2p_{FN,Bound} - 1) + \frac{\sqrt{e^{4\mathcal{I}} - e^{2\mathcal{I}}}}{\sigma_1\sqrt{2}}\Delta\mu_1 - \frac{\sqrt{e^{2\mathcal{I}} - 1}}{\sigma_2\sqrt{2}}\Delta\mu_2}{\sqrt{2e^{4\mathcal{I}} - 3e^{2\mathcal{I}} + 1}} \tag{34}$$

6.4.4 Small Dimension Validation

Fig. 7 shows five plots of $\hat{p}_{FP}$ for different values of displacements $\Delta = \mu_n - \mu_{n,H_0}$. Indeed, the actual value of Δ is unknown. However, the focus is on the minimization of $\hat{p}_{FP}$. Accordingly, $\hat{p}_{FP}$ decreases in each sub-figure of Fig. 7. The actual value of Δ only determines how fast $\hat{p}_{FP}$ decreases. So we can solve the optimization problem (cf. Eq. 1) by minimizing $\hat{p}_{FP}$. We considered two random variables X_1 and X_2, as in our example in Sec. 4 where we created two clusters. A very important factor here is the underlying strategy for clustering test cases. As the distribution of the random variables X_1 and X_2 is directly related to the clustering strategy the main focus is on the maximization of the differential mutual information $\mathcal{I}(X_1; X_2)$. Accordingly, $\mathcal{I}$ is an optimization parameter for effectively reducing $\hat{p}_{FP}$. Lastly, we chose the

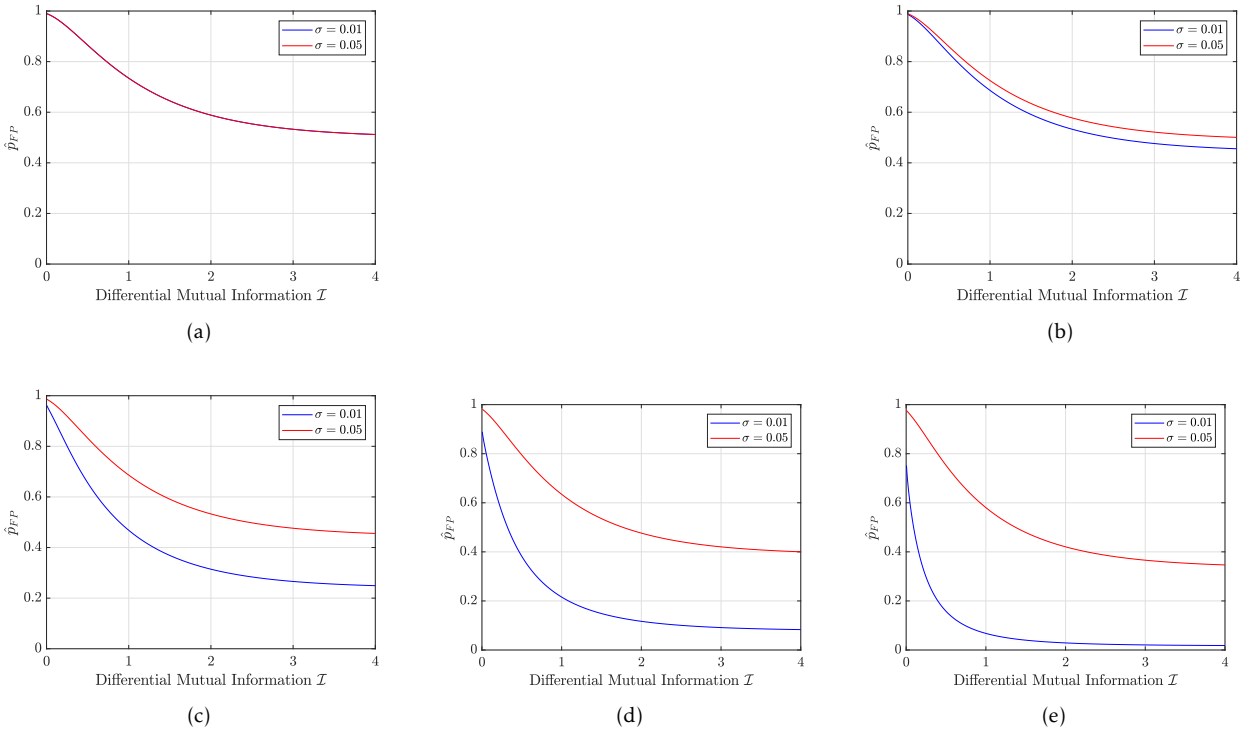


Figure 7: Estimation of $\hat{p}_{FP}$ for different values of $\boldsymbol{\Delta}$ and σ with $\boldsymbol{\Delta} = \alpha \cdot \boldsymbol{v} \cdot 10^{-3}$. The value of α is 0, 1, 5, 10 and 15 in Fig. 7 a), Fig. 7 b), Fig. 7 c), Fig. 7 d) and Fig. 7 e) respectively.

following values $\{0.01, 0.05\}$ for $\sigma = \sigma_1 = \sigma_2$ and selected the following displacements $\boldsymbol{\Delta} = \alpha \cdot \boldsymbol{v} \cdot 10^{-3}$ with $\alpha \in \{0; 1; 5; 10; 15\}$ and $\boldsymbol{v} = [1, -1]^T$.

7 Optimization

The first strategy is to optimize the feature selection. Optimal features are learned in an unsupervised learning session where an evolutionary optimization framework is applied to search for optimal features. The next strategy is to improve the labeling of test cases through an active learning strategy.

7.1 Evolutionary Optimization

Clustering (and sub-clustering) of test cases is performed based on features. Therefore, different clusterings for different selections of feature subsets (Φ_m, Φ_s) are possible. Accordingly, a different statistical model is obtained, as it reflects the failure frequencies in clusters. Furthermore, the differential mutual information (cf. Eq. 31) depends on the statistical dependencies and thus changes for different clusterings.

Sec. 6 proposed a calculation formula for the weights $w_n, 0 \leq n \leq N$, of a linear classifier. However, those formulas still depend on the differential mutual information I. A desired sensitivity has to be guaranteed, and thus the hyperplane is adjusted according to the value of I. It can be shown that for small values of I, the position of the hyperplane still guarantees a desired sensitivity but the false-positive selection probability increases. To minimize the false-positive selection probability, the differential mutual information has to be maximized, which is the final strategy for solving the constrained optimization problem (cf. Eq. 1).

First, clustering depends on the history-tuples of test cases as, for example, the length of the history-tuples determines the maximum number $|S|^K$ of clusters. Second, feature selection is optimized. All in all, we have summarized that K (number of considered previous regressions) is an optimization parameter and $\boldsymbol{b}_s$ (for coding selected and main features) is an optimization matrix. However, this is a large-scale high dimensional optimization problem, as there exist many possible settings for K and $\boldsymbol{b}_s$. Thus, [3] and [4] suggest that the high dimensional optimization problem can be solved in a reasonable time by using evolutionary algorithms. Accordingly, an evolutionary optimization framework is applied for solving the mentioned high dimensional optimization problem. As each setting for K and $\boldsymbol{b}_s$ is one possible solution for clustering test cases, which is the basis for derivation of a stochastic model, the fitness of this solution can be evaluated by calculating the extracted information $\mathcal{I}$ in Eq. 31. Thus, the optimal parameter and matrix setting with the best fitness will survive and will be returned by the evolutionary optimization algorithm.

Fig. 8 shows the overall flow chart of the evolutionary optimization framework. First of all, a new population consisting of several genotypes is initialized. Each genotype stands for a possible setting of K and $\boldsymbol{b}_s$. In the next step, the corresponding phenotypes of the genotypes are derived. Hence each phe-

notype encodes a stochastic model. Accordingly, the population is evaluated, wherein the fitness of each phenotype is calculated. However, a *bad* fitness is also possible due to *bad* statistical properties of the underlying stochastic model. This means that statistical calculations based on the stochastic model that a phenotype encodes cannot guarantee desired statistical confidence bounds. This will be explained in more detail in Sec. 8. Those phenotypes with *bad* fitness cannot survive and hence are eliminated.

Accordingly, remaining genotypes (phenotypes) are stochastically selected, and successively new genotypes are generated due to *crossover* and *mutation* operations. After a certain number of iterations, the phenotype with the best fitness will be selected, and this will be used in the selection algorithm. However, if the population is empty since all phenotypes were of *bad* fitness, then the training mode is activated, in which test cases are still executed without running the selection algorithm.

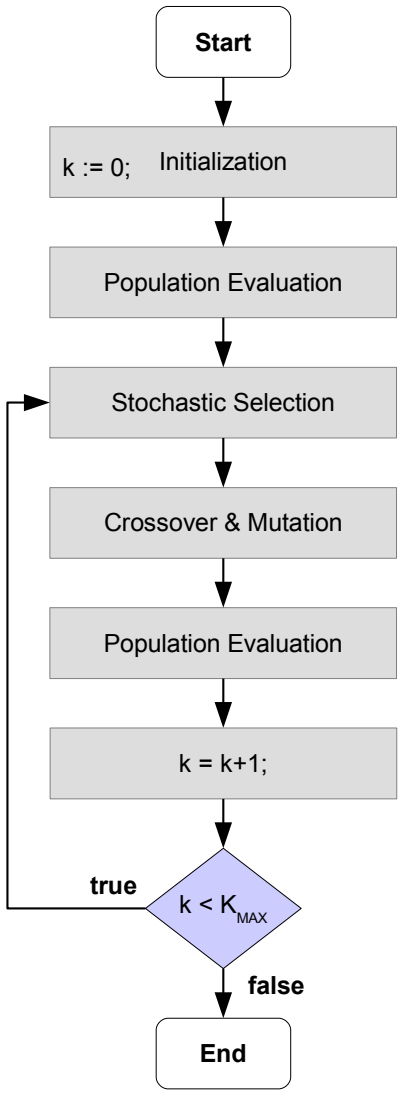


Figure 8: Evolutionary Optimization Framework.

7.2 Active Learning

Our classifier's conducted decisions can be regarded as *hard* or even as *soft* decisions. Once taken, *hard* decisions are never changed later on, in contrast to *soft* decisions. The test efficiency can be significantly increased by conducting *soft* decisions as opposed to *hard* decisions.

7.2.1 Hard Decision

[1] performs *hard decisions* since a selected test case is automatically executed and a once deselected test case is never selected again in the current regression test. In the following passages, the disadvantage of conducting *hard decisions* will be explained in detail in relation to classifier decisions.

The linear classifier's decision depends on the current estimation of $\hat{N}_{FN}$ as it calculates the allowed residual risk $p_{FN,Limit}$ (cf. Eq 6) of potentially taking a wrong decision. $\hat{N}_{FN}$ returns the number of supposedly unrevealed system failures that would be detected by those already deselected test cases that are elements of $T_{\overline{Exec}}$. Accordingly, the linear classifier's decision depends on the decisions it has already taken ($T_{\overline{Exec}}$) and hence it is memory driven.

Each deselected test case t_j has an individual additional contribution $\hat{N}_{FN,j}$ (cf. Eq. 36) to the overall estimation $\hat{N}_{FN}$ such that the relation in Eq. 35 holds.

$$\hat{N}_{FN} = \sum_{\forall t_j \in T_{\overline{Exec}}} \hat{N}_{FN,j} \tag{35}$$

$\hat{N}_{FN,j}$ is the product of t_j's failure probability x and the false-negative probability $P(\hat{H} = H_0|H_1)$ by deselecting t_j as given in Eq. 36.

$$\hat{N}_{FN,j} = P(H_1)P(\hat{H} = H_0|H_1) \tag{36}$$

Because of this fact, a deselection of an arbitrary test case can cause that the residual risk $p_{FN,Limit}$ reaches zero as $\hat{N}_{FN}$ increases (cf. Eq. 6). This means that no more risk ($p_{FN,Limit} = 0$) is allowed, and all remaining test cases have to be consequently selected.

Indeed, selecting test cases even if their deselection is allowed according to risk calculations is sometimes the better choice. In fact, this is the case if $p_{FN,Limit}$ is zero and thus it can be significantly increased by selecting and executing an already deselected test case in order to eliminate its risk. When this is done, $\hat{N}_{FN}$ decreases and hence $p_{FN,Limit}$ increases and thus a residual risk for further deselections is obtained.

However, the amount $\Delta\hat{N}_{FN}$ of how much $\hat{N}_{FN}$ can be decreased by selecting an arbitrary test case is significant. If later more than one test case can be deselected, and these deselected test cases add the same amount of expected unrevealed system failures $\Delta\hat{N}_{FN}$ to $\hat{N}_{FN}$ is in fact a gain in terms of reducing the regression test effort. So the strategy is to deselect primarily those test cases with fewer failure probabilities in order to increase testing efficiency.

As a result, the regression test efficiency can be increased. Therefore, the proposed selection method [1] is extended by a *soft decision* methodology. So each decision for deselecting a test case is now regarded as a *soft decision* that might be changed later. (We note here that the other way round is impossible since an already selected test case is automatically executed on the system under test and hence deselecting it later does not make sense).

7.2.2 Soft Decision

Fig. 9 shows the logic for managing soft selection decisions: Let us assume that t_i is the next test case that is analyzed by the linear classifier. If t_i is deselected, then it is queued into a priority queue whereby its priority is calculated as given in Eq. 37.

$$prio(t_i) = \hat{N}_{FN,i} = P(H_1)P(\hat{H} = H_0|H_1) \qquad (37)$$

In the other case, if t_i is selected then test cases deselected up to this point are analyzed to the end of improving the trade-off between the assumed risk and the total number of deselected test cases. As a consequence, the most probable failing test case t_j is obtained by taking the peek-operation on the priority queue. The priority of t_i and t_j is compared, and the test case with the higher priority is selected and executed on the system under test.

If t_j is executed, then it is removed from the set $T_{\overline{Exec}} \leftarrow T_{\overline{Exec}} \backslash t_j$ and added into the set $T_{Exec} \leftarrow T_{Exec} \cup t_j$. Furthermore, t_j's state is evaluated $eval(t_j)$ and accordingly the empirical failure probabilities of test cases are updated in algorithm 1. Since the calculated failure probabilities are averages of test case evaluations, the failure probabilities of those sub-clusters (see Eq. 4) have to be updated where t_j is an element of them. Accordingly, the failure probabilities of $\forall t_k \in T_{\overline{Exec}}$ are updated in algorithm 1.

Algorithm 1 Test case selection algorithm

```
procedure UPDATE STATISTICS(T_Exec_bar, t_j)        ▷ T_Exec_bar contains already deselected test cases; t_j is executed
    for each test case t_k ∈ T_Exec_bar do
        ∃!C_{n,l} ⟹ t_k ∈ C_{n,l}        ▷ Find sub-cluster of t_k and thus determine n and l
        if t_j ∈ C_{n,l} then
            p_{n,l} ← 1/|C_{n,l}| Σ_{∀t_i∈C_{n,l}} eval(t_i)        ▷ see Eq. 4
            P(H_1) ← p_{n,l}
            update t_k's priority: prio(t_k)        ▷ see Eq. 37
        end if
    end for
    if eval(t_j) == 1 then
        N_TP ← N_TP + 1
    end if
end procedure
```

The important point is that even the failure probability of t_i is computed again. In most cases, t_i would be deselected. Nevertheless, it could be possible that the execution of t_j has failed, such that a further system failure has been found. In such a case, even t_i's failure probability may have increased such that its deselection has to be checked again by the linear classifier.

All in all, testing efficiency can be significantly increased by performing *soft selection* decisions. The performance of both selection strategies (*hard decision* and *soft decision*) will be compared in Sec. 9.

8 Learning Phase

The learning phase is of essential importance due to the fact that during this phase, the system reliability is actually learned. Test case selection is a safety-critical binary classification task as probably system failures would remain undetected and hence, a corresponding quality measure of wrong decisions is required. Accordingly, risk estimations on probably undetected system failures due to deselection of test cases have to be as accurate as possible. The more the system is learned during a regression test, the more precise the risk estimations are. However, learning a system in terms of understanding its reliability is a costly process, as it requires test cases to be executed. The fundamentally important research question is how much training data is enough for safely selecting test cases with a desired sensitivity.

8.1 Statistical Sensitivity Estimation

We have already required a specific sensitivity in the constraint optimization problem (cf. Eq. 1). Accordingly, we define the following confidence level in Eq. 38, which is basically driven from the constraint of Eq. 1.

$$P\Big(\Psi \leq \gamma\Big) \geq 1 - \alpha \qquad (38)$$

Ψ is an estimator for the number of false-negatives $\hat{N}_{FN} = \sum_{t_i \in T_{\overline{EXEC}}} \hat{N}_{FN,i}$ and the bound is given as $\gamma = \frac{N_{TP} \cdot p_{FN,MAX}}{1 - p_{FN,MAX}}$. $\Psi = \sum_i \psi_i$ is composed of several random variables ψ_i standing for the distribution of each $\hat{N}_{FN,i}$. ψ_i's distribution is complex, since the individual contribution of a deselected test case t_i is given by $\hat{N}_{FN,i} = x_N \hat{p}_{FN}$ where x_N is t_i's failure probability and $\hat{p}_{FN}$ is the corresponding estimated false-negative probability: The following theorem is already proved in [1] and gives the formula for the false-negative probability estimation.

Theorem 2 (False-Negative Probability)**.** *For a given p_{th} the calculation formula of the false-negative probability $P(\hat{H} = H_0|H_1)$ has the form*

$$\hat{p}_{FN} = \frac{1}{2}\left(1 + erf\left(\frac{1}{\sqrt{2}} \frac{x_N - (x_N - p_{th})e^{2\mathcal{I}} - \mu_N}{\sigma_N \sqrt{e^{2\mathcal{I}} - 1}}\right)\right) \qquad (39)$$

where x_N is the failure probability of a test case, whereas μ_N, σ_N are parameters of the probability distribution function $\mathcal{N}(\mu_N, \sigma_N)$.

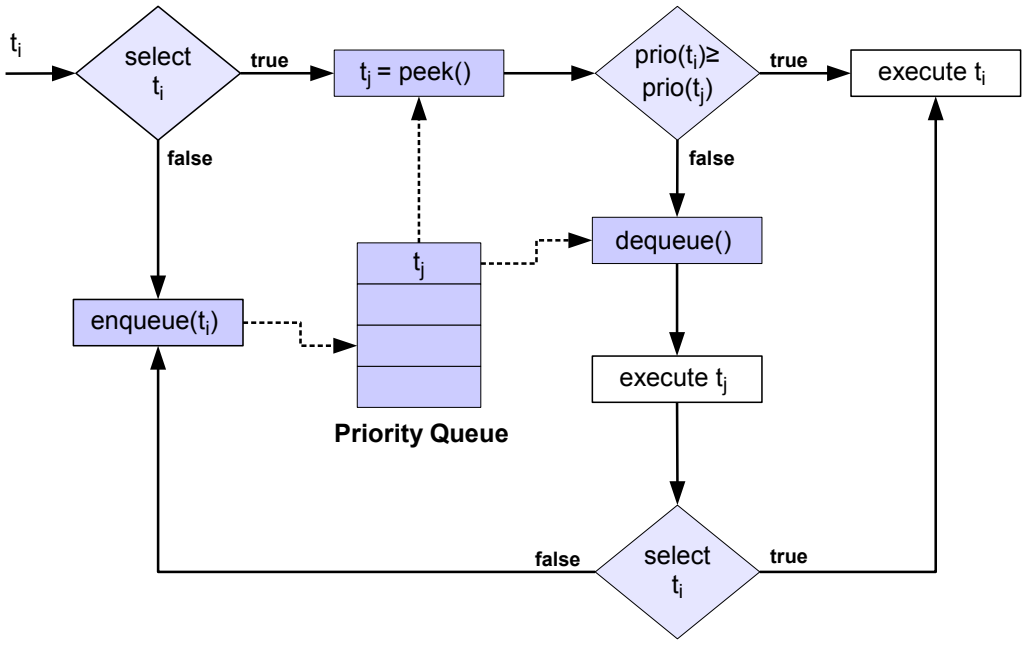


Figure 9: The extended concept for conducting soft decisions

These statistics are created on the basis of sample averages, such that the consideration of sample variances becomes inevitable during the estimation of confidence intervals. For instance, the false-negative probability estimation $\hat{p}_{FN}$ (see Eq. 39) is composed of the statistics x_N, μ_N, σ_N and I. Therefore, its variance depends on the individual variances of each statistic, including the sample variance of x_N. Especially the probability distribution of I is complex as it is nonlinearly composed of a set of multivariate distributed Gaussian random variables.

Therefore, we choose the following approach to solving Eq. 38. We simplify the definition of ψ_i as follows $\psi_i = \hat{p}_{FN} \cdot X_N$ where $\hat{p}_{FN}$ is assumed to be a constant value without any distribution. This step simplifies the calculation complexity of Eq. 38 significantly, as Ψ becomes simply a weighted sum of Gaussian random variables. However, the variance of $\hat{p}_{FN}$ is of course relevant and should not be easily neglected. Accordingly, we require a maximum confidence interval width for $\hat{p}_{FN}$ such that the estimated false-negative probability is quite accurate and hence can be assumed to be just like a constant value without any statistical deviation. We calculate the confidence interval $[\hat{p}_{FN}^{(l)}; \hat{p}_{FN}^{(u)}]$ and its width $\delta = \hat{p}_{FN}^{(u)} - \hat{p}_{FN}^{(l)}$ and require a maximal confidence interval width of δ_{max}.

The Wilson score interval [22] delivers confidence bounds for binomial proportions. Therefore, we calculate the following confidence intervals $[x_n^{(l)}; x_n^{(u)}]$ (confidence level: $1 - \alpha = 99\%$) for each failure probability estimation $x_n, 1 \leq n \leq N$. Each bound is consequently used for building the bounds of the composed statistics σ, Σ and I. By doing this, we obtain the following bounds: $\sigma^{(b)}$, $\Sigma^{(b)}$ and $I^{(b)}$ with $b = \{u, l\}$. Accordingly, we calculate $\hat{p}_{FN}^{(b)}$ by consequently inserting the bounds $\sigma^{(b)}$, $\Sigma^{(b)}$ and $I^{(b)}$ for the statistics σ, Σ and I respectively.

8.2 Criteria for Training

In order to guarantee a statistical bound on the sensitivity with a 99% confidence level, the following conditions have to be checked.

1. $\delta \leq \delta_{max}$
2. $P\left(\Psi \leq \gamma\right) \geq 1 - \alpha = 99\%$

If both conditions are fulfilled, then these risk calculations in the selection algorithm are reasonably accurate and hence selection decisions can be performed. However, if one condition is not fulfilled then the training mode is just active, such that test cases still have to be executed.

9 Industrial Case Study

A German premium car manufacturer constitutes each regression test as being a system release test, and thus the system test takes up to several weeks according to [5]. However, a first detected system failure makes a system release impossible so optimizing the current regression for achieving high efficiency in reducing the regression effort becomes justified.

It is often the case that close to the so-called start of production (SOP) of a vehicle, many electronic control units (ECU) have only some critical spots and thus each regression test is expected to be a system-release test. Since many test cases pass, a lot of time is spent in observing passing test cases. Therefore, reducing the number of executed passed test cases (since a system failure is detected and a system release

is no longer possible) and keeping the limited testing time back for fault-revealing test cases decreases the regression test effort significantly. In any case, a final regression test will succeed after further system updates have been conducted; this will be constituted as a final release-test that meets the high-quality standards of [5].

In our industrial case study, we applied our selection method to a production-ready controller that implements complex networked functionalities for the protection of passengers and other road users. Therefore its test effort is immense, and hence we apply our regression test selection method for accelerating its testing phase. In Fig. 10 the right-hand side of the well-known *V-Model* (see [23]) is shown, whereas the focus is on system testing in our case study.

A hardware-in-the-loop simulator (HiL) [24] is used for validating an ECU's networked complex functionalities as well as its I/O-interaction and its robustness during voltage drops, as it provides an effective platform for testing complex real-time embedded systems.

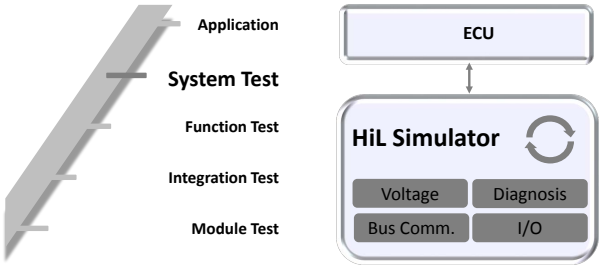


Figure 10: A HiL simulator is used for performing the system test.

Further, we selected for the following test case features for training the machine learning algorithm:

- Name of verified system parts
- Name of a function for which reliability is assured
- Number of totally involved electronic control units during the testing of a networked functionality
- Error type (broken wire etc.) in hardware robustness tests
- Set of checked diagnostic trouble codes, *DTC*s
- Number of checked diagnostic trouble codes, *DTC*s

Since the quality of our selection decisions is hedged on a stochastic level, it can appear that during different runs of our selection method, a statistical deviation of the false-positive probabilities could occur. Therefore, we constitute several independent runs of a regression test, where we set $p_{FN,MAX} = 1\%$. The boxplots and the quantiles of the false-positive probabilities are given in Fig. 11 and in Table 4, respectively.

Fig. 11 shows the overall boxplots of the false-positive probabilities achieved during the regression test replications. To compare the *hard* with the *soft decision* strategy we performed distinct regression test replications where we disabled and enabled the parameter for *'soft decision'*, respectively.

It can be seen from Fig. 11a) and Fig. 11b) that the average false-positive probability is about 74% and 23% for *hard* and enabled *soft decisions* respectively. As already mentioned, conducting *hard decisions* does not allow for global optimization of the trade-off between an already assumed risk and the corresponding number of totally deselected test cases. Global optimization hence requires the analysis of all test cases deselected thus far over and over again, and, if necessary, the selection of an already deselected test case. Therefore, test cases with a higher failure probability should be considered again for eventual selection in an ongoing regression test in order to potentially deselect further less risky test cases. As a result, the regression test effort can be reduced much more by applying *soft decisions*.

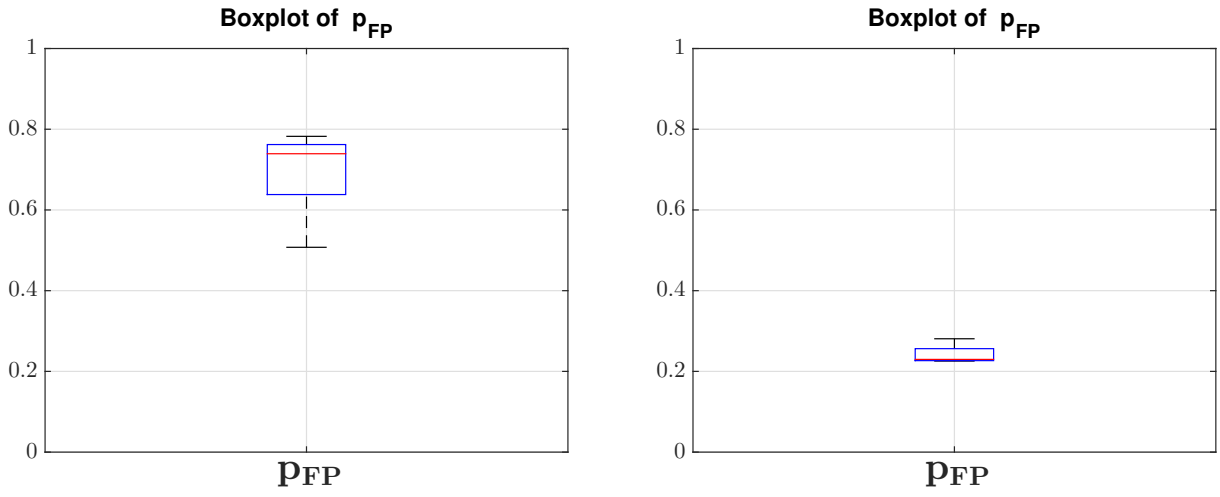
Furthermore, the condition in Eq. 40 on the false-negative probability p_{FN} or on the number of actually occurring false-negatives N_{FN} was fulfilled in all conducted regression test replications.

$$\begin{aligned} &p_{FN} \leq p_{FN,MAX} = 1\% \text{ or} \\ &N_{FN} \leq 1 \end{aligned} \tag{40}$$

Our implemented algorithm for selecting test cases runs on a desktop CPU that is specified in Table 5. We decided to conduct a multithreaded execution of the evolutionary algorithm such that the fitness of all phenotypes in a population is computed in a multithreaded manner (in total 32 threads). Thus the average CPU load is approximately 95% and the maximum memory allocation is about 4GB. We need a mean analysis time of 0.9s for deciding whether a test case should be selected or not.

10 Conclusion and Future Work

We proposed a holistic optimization framework for the safety assessment of systems during regression testing. To this end, we designed a linear classifier for (de-) selecting test cases according to a classification due to a risk-associated recognition. Therefore we defined an optimization problem, since the classifier's specificity has to be maximized whereas its sensitivity still has to exceed a certain threshold $1 - p_{FN,MAX}$. Accordingly, we developed a novel method for determining the weights of a linear classifier that solves the above optimization problem. We have theoretically shown that the classifier performance is directly interrelated with the success of selected relevant features of test cases. Lastly, we applied our method to a production-ready controller and analyzed the overall regression test effort subject to an active learning strategy. We have demonstrated that, in the regression testing of safety-critical systems, significant sav-



(a) hard decision

(b) soft decision

Figure 11: Boxplots of false-positive probabilities in case of a) hard and b) soft decisions

Table 4: Quantiles of the false-positive probabilities in each replicated regression test

	Quantiles of False-Positive Probability				
	0.025	0.25	0.5	0.75	0.975
Hard Decision	50.8%	63.8%	73.9%	76.2%	78.2%
Soft Decision	22.6%	22.7%	23%	25.6%	28.1%

ings can be achieved. As feature selection is a complex task, and thus an evolutionary optimization supposedly finds local optima, more thorough research in this field may indeed allow higher-order reductions of the classifier's false-positive selection probability.

11 Appendix

In the following, a detailed proof of Theorem 1 is given, relating to the proofs given in [1].

Proof of Theorem 1

Proof. According to [1], the maximum likelihood estimation $x_{N,ML}$ (abbreviated x_{ML} in the following) is given in Eq. 41.

$$x_{ML} = \mu_N - \frac{(\boldsymbol{x}_D - \boldsymbol{\mu}_D)^T \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T}{\boldsymbol{\phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T} \tag{41}$$

As $\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T = P_1\Sigma^{-1}P_2^T$ and $\boldsymbol{\phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T = P_2\Sigma^{-1}P_2^T = \left(\Sigma^{-1}\right)_{N,N}$ holds (consult Proof of Theorem 3 in [1]) x_{ML} can be written as given in Eq. 42.

$$x_{ML} =: \sum_{n=1}^{N-1} w_n(x_n - \mu_n) + \mu_N \tag{42}$$

with $w_n = -\frac{\left(\Sigma^{-1}\right)_{n,N}}{\left(\Sigma^{-1}\right)_{N,N}}$. Furthermore, the threshold probability p_{th} was calculated in [1] as given in Eq. 43.

$$p_{th} = x_N + \frac{\sqrt{2}\sigma_N\sqrt{e^{2\mathcal{I}}-1}\,\mathrm{erfinv}(2p_{FN,Bound}-1) + \mu_N - x_N}{e^{2\mathcal{I}}} \tag{43}$$

By introducing the definitions of the weights $w_0 := -\frac{\sqrt{2}\sigma_N\sqrt{e^{2\mathcal{I}}-1}\,\mathrm{erfinv}(2p_{FN,Limit}-1)}{e^{2\mathcal{I}}}$ and $w_N := -\frac{e^{2\mathcal{I}}-1}{e^{2\mathcal{I}}}$ the threshold probability can be written as $p_{TH} = -w_0 + \mu_N - w_N(x_N - \mu_N)$.

Eq. 44 derives the final definition of the hyperplane $y(\boldsymbol{x})$ and the acceptance region of the rival hypothesis H_1 by putting the definitions of x_{ML} and p_{TH} together.

$$\begin{aligned} x_{ML} &\geq p_{TH} \\ \mu_N + \sum_{n=1}^{N-1} w_n(x_n - \mu_n) &\geq -w_0 + \mu_N - w_N(x_N - \mu_N) \\ w_0 + \sum_{n=1}^{N} w_n(x_n - \mu_n) &\geq 0 \\ \underbrace{w_0 + \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu})}_{y(\boldsymbol{x})} &\geq 0 \end{aligned} \tag{44}$$

The estimation of the false-positive probability is

Table 5: Computational/memory requirements and algorithm performance

CPU	Intel Core i7-4800MQ @ 2.7 GHz, 8 Threads
CPU Load	95%
RAM	4 GB
# Allocated Threads	32 Threads
Response Time	0.9 seconds

performed by calculating $\hat{p}_{FP} = P(\hat{H} = H_1|H_0)$. However, as already explained in Sec. 6, the conditional pdf $p(\hat{H}|H)$ is not given and therefore $\hat{p}_{FP}$ cannot be directly estimated. Thus, we assume that the conditional pdf $p(\hat{H}|H)$ is of a Gaussian distribution with mean $E[\boldsymbol{X}|H_0] = \boldsymbol{\mu}_{H_0}$ and covariance matrix $E[(\boldsymbol{X} - \boldsymbol{\mu}_{H_0})(\boldsymbol{X} - \boldsymbol{\mu}_{H_0})^T|H_0] = \Sigma$. However, $\boldsymbol{\mu}_{H_0}$ is an unknown parameter vector; additionally, the second-order moments are assumed to be invariant of an event H_0. We will calculate $\hat{p}_{FP}$ in dependency on the unknown vector $\boldsymbol{\mu}_{H_0}$ and will qualitatively show that $\hat{p}_{FP}$ can be minimized independently of $\boldsymbol{\mu}_{H_0}$ due to an optimization strategy. In showing this, we demonstrate that the concept of our work is validated and mathematically proved.

First of all, Eq. 44 is given in Eq. 45 with an additional term.

$$\begin{aligned} w_0 + \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu}) &\geq 0 \\ w_0 + \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu}_{H_0} - \boldsymbol{\mu}_{H_0}) &\geq 0 \end{aligned} \tag{45}$$

Furthermore, by introducing the definition $\Delta\boldsymbol{\mu} = \boldsymbol{\mu} - \boldsymbol{\mu}_{H_0}$ Eq. 45 can be written as given in Eq. 46.

$$\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{\mu}_{H_0}) \geq -w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu} \tag{46}$$

Accordingly, $\hat{p}_{FP}$ is estimated as given in Eq. 47.

$$\hat{p}_{FP} = P\Big(\boldsymbol{w}^T(\boldsymbol{X} - \boldsymbol{\mu}_{H_0}) \geq -w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu}\Big|H_0\Big) \tag{47}$$

By substituting $U := \sum_{n=1}^{N-1} w_n(X_n - \mu_{n,H_0})$ and $V := w_N(X_N - \mu_{N,H_0})$ with $\mu_{n,H_0} = \left(\boldsymbol{\mu}_{H_0}\right)_n$ the false-positive estimation is given as follows.

$$\hat{p}_{FP} = P\Big(U + V \geq -w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu}\Big|H_0\Big) \tag{48}$$

Based on that fact, that H_0 is given U and V are conditionally independent (see explanation in subsection 6.4.2) of each other and hence the random variable $Z := U + V$ has the mean $E[Z] = E[U] + E[V]$ and the variance $\sigma_Z^2 = \sigma_U^2 + \sigma_V^2$. Since $E[X_n|H_0] = \mu_{n,H_0}$ holds the mean of U and V is zero ($E[U] = 0$; $E[V] = 0$).

The variances of U and V are given in Eq. 49

$$\sigma_U^2 = [w_1 \cdots w_{N-1}]\Sigma_{1,1}[w_1 \cdots w_{N-1}]^T \tag{49}$$

and Eq. 50 respectively.

$$\sigma_V^2 = w_N^2\sigma_N^2 \tag{50}$$

As $[w_1 \cdots w_{N-1}]^T = -\frac{\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T}{\boldsymbol{\phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T}$ holds and by substituting $\beta := \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}\boldsymbol{\phi}^T$ the term $\beta^T\Sigma_{1,1}\beta$ can be simplified in Eq. 51 as already explained in the proof of Theorem 3 in [1].

$$\boldsymbol{\beta}^T\Sigma_{1,1}\boldsymbol{\beta} = -\frac{\det(\Sigma_{1,1})}{\det(\Sigma)} + \sigma_N^2\left[\frac{\det(\Sigma_{1,1})}{\det(\Sigma)}\right]^2 \tag{51}$$

It can easily be seen that the variance of U is equal to $\sigma_U^2 = \frac{\beta^T\Sigma_{1,1}\beta}{\left[\left(\Sigma^{-1}\right)_{N,N}\right]^2}$. Furthermore, $\left(\Sigma^{-1}\right)_{N,N} = \frac{\det(\Sigma_{1,1})}{\det(\Sigma)}$ holds and thus the variance is further simplified and is given in Eq. 52.

$$\sigma_U^2 = -\frac{\det(\Sigma)}{\det(\Sigma_{1,1})} + \sigma_N^2 \tag{52}$$

Finally, the false-positive probability is estimated as follows:

$$\hat{p}_{FP} = P(Z \geq -w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu}) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{-w_0 + \boldsymbol{w}^T\Delta\boldsymbol{\mu}}{\sqrt{2}\sigma_Z}\right)\right] \tag{53}$$

For the case $N = 2$, Eq. 53 can be simplified and after several calculation steps the following Eq. 54 results

$$\hat{p}_{FP} = \frac{1}{2}\left[1 - \mathrm{erf}\Big(\psi\Big)\right] \tag{54}$$

with

$$\psi = \frac{\sqrt{e^{2\mathcal{I}} - 1}\,\mathrm{erfinv}(2p_{FN,Bound} - 1) + \frac{\sqrt{e^{4\mathcal{I}} - e^{2\mathcal{I}}}}{\sqrt{2}\sigma_1}\Delta\mu_1 - \frac{\sqrt{e^{2\mathcal{I}} - 1}}{\sqrt{2}\sigma_2}\Delta\mu_2}{\sqrt{2e^{4\mathcal{I}} - 3e^{2\mathcal{I}} + 1}} \tag{55}$$

□

References

[1] I. Alagöz, T. Herpel, and R. German, "A selection method for black box regression testing with a statistically defined quality level," in *2017 IEEE International Conference on Software Testing, Verification and Validation (ICST)*, March 2017, pp. 114–125, DOI: 10.1109/ICST.2017.18.

[2] F. Bürger and J. Pauli, "Representation optimization with feature selection and manifold learning in a holistic classification framework," in *Proceedings of the International Conference on Pattern Recognition Applications and Methods*, 2015, pp. 35–44, DOI: 10.5220/0005183600350044.

[3] T. Back, *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press, 1996, DOI: 10.1108/k.1998.27.8.979.4.

[4] F. Bürger and J. Pauli, "Understanding the interplay of simultaneous model selection and representation optimization for classification tasks," in *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods*, 2016, pp. 283–290, DOI: 10.5220/0005705302830290.

[5] "ISO/DIS 26262-10 - Road vehicles — Functional safety," http://www.iso.org, Tech. Rep., 2012, DOI: doi:10.3403/30205385.

[6] G. Xie and Z. Dang, "Model-checking driven black-box testing algorithms for systems with unspecified components," *CoRR*, vol. cs.SE/0404037, 2004. [Online]. Available: http://arxiv.org/abs/cs.SE/0404037

[7] R. Grosu and S. A. Smolka, *Monte Carlo Model Checking*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 271–286, DOI: 10.1007/978-3-540-31980-1_18. [Online]. Available: https://doi.org/10.1007/978-3-540-31980-1_18

[8] A. Legay, B. Delahaye, and S. Bensalem, "Statistical model checking: An overview," in *International Conference on Runtime Verification*. Springer, 2010, pp. 122–135, DOI: 10.1007/978-3-642-16612-9_11.

[9] E. Elkind, B. Genest, D. Peled, and H. Qu, *Grey-Box Checking*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 420–435, DOI: 10.1007/11888116_30. [Online]. Available: https://doi.org/10.1007/11888116_30

[10] K. Sen, M. Viswanathan, and G. Agha, "Statistical model checking of black-box probabilistic systems," in *International Conference on Computer Aided Verification*. Springer, 2004, pp. 202–215, DOI: 10.1007/978-3-540-27813-9_16.

[11] D. Peled, M. Y. Vardi, and M. Yannakakis, *Black Box Checking*. Boston, MA: Springer US, 1999, pp. 225–240, DOI: 10.1007/978-0-387-35578-8_13. [Online]. Available: https://doi.org/10.1007/978-0-387-35578-8_13

[12] S. Yoo and M. Harman, "Regression testing minimization, selection and prioritization: a survey," *Software Testing, Verification and Reliability*, vol. 22, no. 2, pp. 67–120, 2012, DOI: 10.1002/stvr.430. [Online]. Available: http://dx.doi.org/10.1002/stvr.430

[13] G. Rothermel and M. J. Harrold, "A safe, efficient regression test selection technique," *ACM Trans. Softw. Eng. Methodol.*, vol. 6, no. 2, pp. 173–210, Apr. 1997, DOI: 10.1145/248233.248262. [Online]. Available: http://doi.acm.org/10.1145/248233.248262

[14] A. Orso, M. J. Harrold, D. Rosenblum, G. Rothermel, M. L. Soffa, and H. Do, "Using component metacontent to support the regression testing of component-based software," in *Proceedings IEEE International Conference on Software Maintenance. ICSM 2001*, 2001, pp. 716–725, DOI: 10.1109/ICSM.2001.972790.

[15] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, R. Anil, Z. Haque, L. Hong, V. Jain, X. Liu, and H. Shah, "Wide & deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, ser. DLRS 2016, vol. abs/1606.07792. New York, NY, USA: ACM, 2016, pp. 7–10, DOI: 10.1145/2988450.2988454. [Online]. Available: http://doi.acm.org/10.1145/2988450.2988454

[16] R. Langone, O. M. Agudelo, B. De Moor, and J. A. Suykens, "Incremental kernel spectral clustering for online learning of non-stationary data," *Neurocomputing*, vol. 139, pp. 246–260, 2014, DOI: 10.1016/j.neucom.2014.02.036 .

[17] S. Mehrkanoon, O. M. Agudelo, and J. A. Suykens, "Incremental multi-class semi-supervised clustering regularized by kalman filtering," *Neural Networks*, vol. 71, pp. 88–104, 2015, DOI: 10.1016/j.neunet.2015.08.001.

[18] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009. [Online]. Available: http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf

[19] C. Lin, M. Mausam, and D. Weld, "Re-active learning: Active learning with relabeling," 2016. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12500

[20] F. Bürger and J. Pauli, "Automatic representation and classifier optimization for image-based object recognition," in *VISAPP 2015 - Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Volume 2, Berlin, Germany, 11-14 March, 2015.*, 2015, pp. 542–550, DOI: 10.5220/0005359005420550. [Online]. Available: https://doi.org/10.5220/0005359005420550

[21] *Handbook of Mathematics*. Springer Berlin Heidelberg, 2007, DOI: 10.1007/978-3-540-72122-2. [Online]. Available: https://doi.org/10.1007%2F978-3-540-72122-2

[22] M. Thulin, "The cost of using exact confidence intervals for a binomial proportion," *Electron. J. Statist.*, vol. 8, no. 1, pp. 817–840, 2014, DOI: 10.1214/14-EJS909. [Online]. Available: https://doi.org/10.1214/14-EJS909

[23] I. Sommerville, *Software Engineering: (Update) (9th Edition) (International Computer Science)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2011.

[24] M. Schlager, *Hardware-in-the-Loop Simulation: A Scalable, Component-based, Time-triggered Hardware-in-the-loop Simulation Framework*. VDM Verlag Dr. Müller E.K., 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 66-75 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Numerical Solution of Fuzzy Differential Equations with Z-numbers using Fuzzy Sumudu Transforms

Sina Razvarz[1], Raheleh Jafari[*,2], Wen Yu[1]

[1]*Departamento de Control Automatico CINVESTAV-IPN (National Polytechnic Institute) Mexico City 07360, Mexico*

[2]*Department of Information and Communication Technology Agder University College, 4876 Grimstad, Norway*

ARTICLE INFO

Article history:
Received: 14 November, 2017
Accepted: 17 January, 2017
Online: 30 January, 2018

Keywords:
K fuzzy Sumudu transform
fuzzy differential equation
Z-number

ABSTRACT

The uncertain nonlinear systems can be modeled with fuzzy differential equations (FDEs) and the solutions of these equations are applied to analyze many engineering problems. However, it is very difficult to obtain solutions of FDEs.
In this paper, the solutions of FDEs are approximated by utilizing the fuzzy Sumudu transform (FST) method. Here, the uncertainties are in the sense of Z-numbers. Important theorems are laid down to illustrate the properties of FST. The theoretical analysis and simulation results show that this new technique is effective to estimate the solutions of FDEs.

1. Introduction

This paper is an extension of work originally presented in [1]. In many physical and dynamical processes, mathematical modeling leads to the deterministic initial and boundary value problems. In practical the boundary values may be different from crisp and displays in the form of unknown parameters [2]. When the parameters or the states of the differential equations are uncertain, they can be modeled with FDE. In recent days, many methods have used FDE for modeling and control of uncertain nonlinear systems [3-5]. The basic idea of the fuzzy derivative was first introduced in [Chang]. Then it is extended in [6]. The first-order fuzzy initial value problem, as well as fuzzy partial differential equation, have been studied in [7]. By generalizing the differentiability, [6] gave an analytical solution. The Lipschitz condition, as well as the theorem for existence and uniqueness of the solution related to FDEs, are discussed in [10-12]. In [13], the analytical solutions of second order FDE are obtained. The analytical solutions of third order linear FDE are found in [14]. By the interval-valued method, [15] examined the basic solutions of nonlinear FDEs with generalized differentiability.

A novel technique in order to solve FDEs is laid down based on the Sumudu transform. Sumudu transform along with broad applications has been utilized in the area of system engineering and applied physics [16-18]. In [19], some simple and deeper fundamental theorems, as well as properties of the Sumudu Transform, were generalized. In [20], Sumudu transform is applied to the system of differential equations. In [21], Sumudu transform is used in order to find the solution of the fuzzy partial differential equation. In [22], Sumudu transform has been used to solve fractional differential equations.

In this paper, we use FST to approximate the Z-number solutions of the FDEs. The FST reduces the FDE to an algebraic equation. A very important property of the FST is that it can solve the equation without resorting to a new frequency domain. The procedure of switching FDEs to an algebraic equation is cited in [10] and is stated as an operational calculus. We extend our previous work [1] by generating more theorems for describing the properties of FST and displaying the uncertainties with Z-numbers. The Z-number is a new concept that is subjected to a higher potential to demonstrate the information of the human being as well as to utilize in information processing [23]. Z-numbers can be regarded as to answer questions and carry out the decisions [24]. There exist few structure based on the theoretical concept of Z-numbers [25]. [26] gave an inception, which results in the extension of the Z-numbers. [27] generated a theorem to convert the Z-numbers to the usual fuzzy sets.

In this paper, initially, some preliminary definitions along with properties related to FST are demonstrated. After that, solving FDEs by using the methodology of FST has been discussed. At the end, two examples along with comparisons are utilized in order to demonstrate the effectiveness of our proposed method.

*Corresponding Author: Raheleh Jafari, jafari3339@yahoo.com

2. Preliminaries

Prior to the introduction of the FST, some concepts related to the fuzzy variables and Z-numbers are laid down in this section [28, 29].

Definition 1: A fuzzy number B is a function of $B \in E : R \to [0,1]$, in such a manner, 1) B is normal, (there exists $a_0 \in R$ in such a manner $B(a_0) = 1$; 2) B is convex, $B(\gamma a + (1-\gamma)c) \geq \min\{B(a), B(c)\}$, $\forall a, c \in R, \forall \gamma \in [0,1]$; 3) B is upper semi-continuous on R, i.e., $B(a) \leq B(a_0) + \varepsilon$, $\forall a \in N(a_0)$, $\forall a_0 \in R$, $\forall \varepsilon > 0$, $N(a_0)$ is a neighborhood; 4) The set $B^+ = \{a \in R, B(a) > 0\}$ is compact.

Definition 2: The r-level of the fuzzy number B is defined as follows

$$[B]^r = \{a \in R : B(a) \geq r\} \tag{1}$$

where $0 < r \leq 1$, $B \in E$.

Definition 3: Let $B_1, B_2 \in E$ and $\xi \in R$, the operations addition, subtraction, multiplication and scalar multiplication are defined as

$$[B_1 \oplus B_2]^r = [B_1]^r + [B_2]^r = [\underline{B}_1^r + \underline{B}_2^r, \overline{B}_1^r + \overline{B}_2^r] \tag{2}$$

$$[B_1 ? B_2]^r = [B_1]^r - [B_2]^r = [\underline{B}_1^r - \underline{B}_2^r, \overline{B}_1^r - \overline{B}_2^r] \tag{3}$$

$$[B_1 ? B_2]^r = \begin{pmatrix} \min\{\underline{B}_1^r \underline{B}_2^r, \underline{B}_1^r \overline{B}_2^r, \overline{B}_1^r \underline{B}_2^r, \overline{B}_1^r \overline{B}_2^r\} \\ \max\{\underline{B}_1^r \underline{B}_2^r, \underline{B}_1^r \overline{B}_2^r, \overline{B}_1^r \underline{B}_2^r, \overline{B}_1^r \overline{B}_2^r\} \end{pmatrix} \tag{4}$$

$$[\xi B_1]^r = \xi [B_1]^r = \begin{cases} (\xi \underline{B}_1^r, \xi \overline{B}_1^r), & \xi \geq 0 \\ (\xi \overline{B}_1^r, \xi \underline{B}_1^r), & \xi \leq 0 \end{cases} \tag{5}$$

Definition 4: The Hausdorff distance between two fuzzy numbers B_1 and B_2 is defined as [30,31]

$$D(B_1, B_2) = \sup_{0 \leq r \leq 1} \{\max(|\underline{B}_1^r - \underline{B}_2^r|, |\overline{B}_1^r - \overline{B}_2^r|)\} \tag{6}$$

$D(B_1, B_2)$ has the following properties

(i) $D(B_1 \oplus u, B_2 \oplus u) = D(B_1, B_2)$, $\forall B_1, B_2, u \in E$

(ii) $D(\xi B_1, \xi B_2) = |\xi| D(B_1, B_2)$, $\forall \xi \in R, B_1, B_2 \in E$

(iii) $D(B_1 \oplus B_2, u \oplus v) \leq D(B_1, u) + D(B_2, v)$, $\forall B_1, B_2, u, v \in E$

(iv) (D, E) is stated as complete metric space.

Definition 5: The function $\psi : [a_1, a_2] \to E$ is integrable on $[a_1, a_2]$, if it satisfies in the below mentioned relation

$$\int_{a_1}^{\infty} \psi(x)dx = \left(\int_{a_1}^{\infty} \underline{\psi}(x,r)dx, \int_{a_1}^{\infty} \overline{\psi}(x,r)dx\right) \tag{7}$$

If $\psi(x)$ be a fuzzy value function, as well as $q(x)$ be a fuzzy Riemann integrable on $[a_1, \infty]$ so $\psi(x) \oplus q(x)$ can be a fuzzy Riemann integrable on $[a_1, \infty]$. Therefore,

$$\int_{a_1}^{\infty} (\psi(x) \oplus q(x))dx = \int_{a}^{\infty} \psi(x)dx \oplus \int_{a}^{\infty} q(x)dx \tag{8}$$

According to the fuzzy concept or in the case of interval arithmetic, equation $B_1 = B_2 \oplus s$ is not equivalent with $s = B_1 ? B_2 = B_1 \oplus (-1)B_2$ or to $B_2 = B_1 ? s = B_1 \oplus (-1)s$ and this is the main reason in introducing the following Hukuhara difference (H-difference).

Definition 6: The definition of H-difference [32,9], is proposed by $B_1 ?_H B_2 = s \Leftrightarrow B_1 = B_2 \oplus s$. If $B_1 ?_H B_2$ prevails, its r-level is $[B_1 ?_H B_2]^r = [\underline{B}_1^r - \underline{B}_2^r, \overline{B}_1^r - \overline{B}_2^r]$.

Precisely, $B_1 ?_H B_1 = 0$ but $B_1 ? B_1 \neq 0$.

Definition 7: Suppose $\psi : [a_1, a_2] \to E$ and $x_0 = [a_1, a_2]$. ψ is strongly generalized differentiable at x_0, if for all $k > 0$ adequately minute, $\psi'(x_0) \in E$ exists in such a manner that (i) $\exists$ $\psi(x_0 + k) ?_H \psi(x_0)$, $\psi(x_0) ?_H \psi(x_0 - k)$ and

$$\lim_{k \to 0^+} \frac{\psi(x_0+k) ?_H \psi(x_0)}{k} = \lim_{k \to 0^+} \frac{\psi(x_0) ?_H \psi(x_0-k)}{k} = \psi'(x_0)$$

or (ii) $\exists\ \psi(x_0) ?_H \psi(x_0 + k)$, $\psi(x_0 - k) ?_H \psi(x_0)$ and

$$\lim_{k \to 0^+} \frac{\psi(x_0) ?_H \psi(x_0+k)}{(-k)} = \lim_{k \to 0^+} \frac{\psi(x_0-k) ?_H \psi(x_0)}{(-k)} = \psi'(x_0),$$

or (iii) $\exists\ \psi(x_0 + k) ?_H \psi(x_0)$, $\psi(x_0 - k) ?_H \psi(x_0)$ and

$$\lim_{k \to 0^+} \frac{\psi(x_0+k) ?_H \psi(x_0)}{k} = \lim_{k \to 0^+} \frac{\psi(x_0-k) ?_H \psi(x_0)}{(-k)} = \psi'(x_0)$$

or (iv) $\exists\ \psi(x_0) ?_H \psi(x_0 + k)$, $\psi(x_0) ?_H \psi(x_0 - k)$ and

$$\lim_{k \to 0^+} \frac{\psi(x_0) ?_H \psi(x_0+k)}{(-k)} = \lim_{k \to 0^+} \frac{\psi(x_0) ?_H \psi(x_0-k)}{k} = \psi'(x_0)$$

Remark 1:It is clear that case (i) is H-derivative. Furthermore, a function is (i)-differentiable only when it is H-derivative. *Remark 2*: It can be concluded from [32] that, the definition of differentiability is noncontradictory [33].

Let us consider $\psi : R \to E$ where $\psi(t)$ has a parametric form as $[\psi(t,r)] = [\underline{\psi}(t,r), \overline{\psi}(t,r)]$, for all $0 \le r \le 1$, thus (i) If ψ be (i)-differentiable, so $\underline{\psi}(t,r)$ and $\overline{\psi}(t,r)$ are differentiable functions, moreover $\psi'(t) = (\underline{\psi}'(t,r), \overline{\psi}'(t,r))$. (ii) If ψ be (ii)-differentiable, so $\underline{\psi}(t,r)$ and $\overline{\psi}(t,r)$ are differentiable functions, moreover $\psi'(t) = (\overline{\psi}'(t,r), \underline{\psi}'(t,r))$. Suppose $f : (a_1, a_2) \to R$ is differentiable on (a_1, a_2), furthermore ψ' has finite root in (a_1, a_2), and $m \in E$, therefore, $\psi(x) = mf(x)$ is strongly generalized differentiable on (a_1, a_2) along with $\psi'(x) = mf'(x)$, $\forall x \in (a_1, a_2)$.

Theorem 1: [9] Assume $\psi : R \times E \to E$ is taken to be a continuous fuzzy function. If $x_0 \in R$, the fuzzy initial value constraint

$$\begin{cases} \phi'(t) = \psi(x, \phi) \\ \phi(x_0) = \phi_0 \end{cases} \tag{9}$$

is incorporated with two solutions: (i)-differentiable, also (ii)-differentiable. Hence the successive iterations

$$\phi_{n+1}(x) = \phi_0 + \int_{x_0}^{x} \psi(t, \phi_n(t))dt, \qquad \forall x \in [x_0, x_1] \tag{10}$$

and

$$\varphi_{n+1}(x) = \varphi_0 \,?_H (-1) \int_{x_0}^{x} \psi(t, \varphi_n(t))dt, \qquad \forall x \in [x_0, x_1] \tag{11}$$

approaches towards the two solutions sequentially.

Definition 8: A Z-number has two components $Z = [B(a), \tilde{p}]$. The primary component $B(a)$ is restriction on a real-valued uncertain variable a. The secondary component $\tilde{p}$ is a measure of the reliability of B. $\tilde{p}$ can be reliability, the strength of belief, probability or possibility. When $B(a)$ is a fuzzy number and $\tilde{p}$ is the probability distribution of a, the Z-number is stated as Z^+-number. When $B(a)$ as well as $\tilde{p}$ are fuzzy numbers, the Z-number is stated as Z^- -number.

The Z^+ -number carries more information when compared with Z^- -number. In this paper, we utilize the definition of Z^+ -number, i.e., $Z = [B, \tilde{p}]$, B is a fuzzy number and $\tilde{p}$ is a probability distribution.

To express the fuzzy number the most common membership functions are utilized in this paper. The popular membership functions are the triangular function

$$\mu_B = F(\lambda_1, \lambda_2, \lambda_3) = \begin{cases} \frac{a-\lambda_1}{\lambda_2-\lambda_1} & \lambda_1 \le a \le \lambda_2 \\ \frac{\lambda_3-\zeta}{\lambda_3-\lambda_2} & \lambda_2 \le a \le \lambda_3 \end{cases} otherwise \mu_B = 0 \tag{12}$$

and trapezoidal function

$$\mu_B = F(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = \begin{cases} \frac{a-\lambda_1}{\lambda_2-\lambda_1} & \lambda_1 \le a \le \lambda_2 \\ \frac{\lambda_4-a}{\lambda_4-\lambda_3} & \lambda_3 \le a \le \lambda_4 \\ 1 & \lambda_2 \le a \le \lambda_3 \end{cases} otherwise \mu_B = 0 \tag{13}$$

The probability measure is defined as

$$\tilde{P} = \int_R \mu_B(a)\tilde{p}(a)da \tag{14}$$

where $\tilde{p}$ is the probability density of a, also R is the restriction on $\tilde{p}$. For discrete Z-numbers

$$\tilde{P}(B) = \sum_{i=1}^{n} \mu_B(a_i)\tilde{p}(a_i) \tag{15}$$

Definition 9: The r -level of the Z-number $Z = (B, \tilde{P})$ is illustrated as

$$[Z]^r = \left([B]^r, [\tilde{p}]^r\right) \tag{16}$$

where $0 < r \le 1$. $[\tilde{p}]^r$ is computed by the Nguyen's theorem

$$[\tilde{p}]^r = \tilde{p}([B]^r) = \tilde{p}([\underline{B}^r, \overline{B}^r]) = \left[\underline{\tilde{P}}^r, \overline{\tilde{P}}^r\right] \tag{17}$$

where $\tilde{p}([B]^r) = \{\tilde{p}(a) \mid a \in [B]^r\}$. So $[Z]^r$ can be demonstrated as the form r -level of a fuzzy number

$$[Z]^r = \left(\underline{Z}^r, \overline{Z}^r\right) = \left(\left(\underline{B}^r, \underline{\tilde{P}}^r\right), \left(\overline{B}^r, \overline{\tilde{P}}^r\right)\right) \tag{18}$$

where $\underline{\tilde{P}}^r = \underline{B}^r \tilde{p}(\underline{a_i}^r)$, $\overline{\tilde{P}}^r = \overline{B}^r \tilde{p}(\overline{a_i}^r)$, $[a_i]^r = (\underline{a_i}^r, \overline{a_i}^r)$.

Similar with the fuzzy numbers, the Z-numbers are also incorporated with three primary operations, addition, subtraction, and multiplication. The operations in this paper are different definitions with [34]. The r -level of Z-numbers is applied to simplify the operations.

Suppose $Z_1 = (B_1, \tilde{p}_1)$ and $Z_2 = (B_2, \tilde{p}_2)$ be two discrete Z-numbers expressing the uncertain variables a_1 and a_2, $\sum_{\iota=1}^{n} \tilde{p}_1(a_{1\iota}) = 1$, $\sum_{\iota=1}^{n} \tilde{p}_2(a_{2\iota}) = 1$. The operations are displayed as

$$Z_{12} = Z_1 * Z_2 = (B_1 * B_2, \tilde{p}_1 * \tilde{p}_2) \tag{19}$$

where $* \in \{\oplus, ?, ?\}$.

For all $\tilde{p}_1 * \tilde{p}_2$ operations, we use convolutions for the discrete probability distributions

$$\tilde{p}_1 * \tilde{p}_2 = \sum_i \tilde{p}_1(a_{1,i})\tilde{p}_2(a_{2,(n-i)}) = \tilde{p}_{12}(a) \tag{20}$$

The above definitions satisfy the Hukuhara difference [35],

$$\begin{aligned} Z_1 \,?_H Z_2 &= Z_{12} \\ Z_1 &= Z_2 \oplus Z_{12} \end{aligned} \tag{21}$$

If $Z_1 ?_H Z_2$ prevails, the r -level is

$$[Z_1 ?_H Z_2]^r = [\underline{Z}_1^r - \underline{Z}_2^r, \overline{Z}_1^r - \overline{Z}_2^r] \quad (22)$$

Obviously, $Z_1 ?_H Z_1 = 0$, $Z_1 ? Z_1 \neq 0$.

Also the above definitions satisfy the generalized Hukuhara difference [7]

$$Z_1 ?_{gH} Z_2 = Z_{12} \Leftrightarrow \begin{cases} 1)\ Z_1 = Z_2 \oplus Z_{12} \\ 2)\ Z_2 = Z_1 \oplus (-1)Z_{12} \end{cases} \quad (23)$$

It is easy to display that 1) and 2) in combination are genuine if and only if Z_{12} is a crisp number. With respect to r -level we have

$[Z_1 ?_{gH} Z_2]^r = [\min\{\underline{Z}_1^r - \underline{Z}_2^r, \overline{Z}_1^r - \overline{Z}_2^r\}, \max\{\underline{Z}_1^r - \underline{Z}_2^r, \overline{Z}_1^r - \overline{Z}_2^r\}]$

and If and subsist, $Z_1 ?_H Z_2 = Z_1 ?_{gH} Z_2$. The circumstances for the inerrancy of $Z_{12} = Z_1 ?_{gH} Z_2 \in E$ are

$$1)\begin{cases} \underline{Z}_{12}^r = \underline{Z}_1^r - \underline{Z}_2^r and \overline{Z}_{12}^r = \overline{Z}_1^r - \overline{Z}_2^r \\ with\ \underline{Z}_{12}^r increa\sin g,\ \overline{Z}_{12}^r decrea\sin g,\ \underline{Z}_{12}^r \leq \overline{Z}_{12}^r \end{cases}$$
$$2)\begin{cases} \underline{Z}_{12}^r = \overline{Z}_1^r - \overline{Z}_2^r and\ \overline{Z}_{12}^r = \underline{Z}_1^r - \underline{Z}_2^r \\ with\ \underline{Z}_{12}^r increa\sin g,\ \overline{Z}_{12}^r decrea\sin g,\ \underline{Z}_{12}^r \leq \overline{Z}_{12}^r \end{cases} \quad (24)$$

where $\forall r \in [0,1]$

If B is a triangular function, the absolute value of the Z-number $Z = (B, \tilde{p})$ is defined as

$$|Z(a)| = (|\lambda_{11}| + |\lambda_{12}| + |\lambda_{13}|, \tilde{p}(|\lambda_{21}| + |\lambda_{22}| + |\lambda_{23}|)) \quad (25)$$

If B_1, as well as B_2 are triangular functions, the supremum metric for Z-numbers $Z_1 = (B_1, \tilde{p}_1)$ and $Z_2 = (B_2, \tilde{p}_2)$ is illustrated as

$$D(Z_1, Z_2) = d(B_1, B_2) + d(\tilde{p}_1, \tilde{p}_2) \quad (26)$$

in this case $d(\cdot,\cdot)$ is defined as the supremum metrics with fuzzy sets [3]. $D(Z_1, Z_2)$ is incorporated with the following possessions

$$D(Z_1 + Z, Z_2 + Z) = D(Z_1, Z_2)$$
$$D(Z_2, Z_1) = D(Z_1, Z_2)$$
$$D(\zeta Z_1, \zeta Z_2) = |\zeta| D(Z_1, Z_2)$$
$$D(Z_1, Z_2) \leq D(Z_1, Z) + D(Z, Z_2)$$

where $\zeta \in R$, $Z = (B, \tilde{p})$ is Z-number and B is triangle function, for proof refer to [36].

Definition 10: Suppose $\hat{Z}$ demonstrates the space of Z -numbers, then the $r-$ level of Z -number valued function $\Psi : [a_1, a_2] \to \hat{Z}$ is defined as

$$\Psi(\phi, r) = [\underline{\Psi}(\phi, r), \overline{\Psi}(\phi, r)]$$

where $\phi \in \hat{Z}$, and $r \in [0,1]$.

Based on the definition of Generalized Hukuhara difference, the gH-derivative of Ψ at ϕ_0 is defined as

$$\Psi'(\varphi_0) = \lim_{h \to 0} \frac{1}{h} [\Psi(\varphi_0 + h) ?_{gH} \Psi(\varphi_0)] \quad (27)$$

In (27), $\Psi(\phi_0 + h)$ and $\Psi(\phi_0)$ represents similar style with Z_1 and Z_2respectively defined in (laiop).

By implementing the $r-$ level (16) to initial value problem, $\phi'(t) = \psi(t, \phi(t))$, we generate two Z-number valued functions: $\underline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)]$ and $\overline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)]$.

The initial value problem can be equivalent to the following relation

$$i)\begin{cases} \underline{\phi}' = \underline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)] \\ \overline{\phi}' = \overline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)] \end{cases}$$
$$ii)\begin{cases} \underline{\phi}' = \overline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)] \\ \overline{\phi}' = \underline{\psi}[t, \underline{\phi}(a,r), \overline{\phi}(a,r)] \end{cases} \quad (28)$$

3. Fuzzy Sumudu transform

Fuzzy initial and boundary value problems can be resolved by utilizing fuzzy Laplace transform [10]. In this paper, the FST methodology for Z-number is illustrated, furthermore the properties of this methodology is stated. By applying the FST methodology, the FDE based on Z-number is reduced to an algebraic equation. The main advantage of the FST is that it can resolve the equation without resorting to a new frequency domain. The methodology of converting FDEs to an algebraic equation is expressed in [10].

Definition 11: Suppose $\psi(t)$ be a continuous Z-number valued function, also, $\psi(Bt) ? e^{-t}$ be an improper Z-number Riemann integrable on $[0, \infty)$. Accordingly, $\int_0^\infty \psi(Bt) ? e^{-t} dt$ is expressed as FST and it is defined by $\Omega(B) = \mathbf{S}[\psi(t)] = \int_0^\infty \psi(Bt) ? e^{-t} dt$, where $0 \leq B < K$, $K \geq 0$, also e^{-t} is a real-valued function. Based on the Theorem 3 we have the following relation

$$\int_0^\infty \psi(Bt) ? e^{-t} dt = (\int_0^\infty \underline{\psi}(Bt, r) e^{-t} dt, \int_0^\infty \overline{\psi}(Bt, r) e^{-t} dt) \quad (29)$$

Let

$$\mathbf{S}[\underline{\psi}(t,r)] = \int_0^\infty \underline{\psi}(Bt, r) e^{-t} dt$$
$$\mathbf{S}[\overline{\psi}(t,r)] = \int_0^\infty \overline{\psi}(Bt, r) e^{-t} dt \quad (30)$$

hence we obtain the following relation

$$\mathbf{S}[\psi(t)] = (\mathbf{S}[\underline{\psi}(t,r), \mathbf{S}\overline{\psi}(t,r)]) \quad (31)$$

Theorem 2: Suppose $\psi'(t)$ be a Z-number valued integrable function, as well as $\psi(t)$ be the primitive of $\psi'(t)$ on $[0,\infty)$. Therefore,

$$\mathbf{S}[\psi'(t)]=\frac{1}{B}?\mathbf{S}[\psi(t)]?(\frac{1}{B}?[\psi(0)]) \tag{32}$$

where is considered to be (i)-differentiable, or

$$\mathbf{S}[\psi'(t)]=\frac{-1}{B}?[\psi(0)]?(\frac{-1}{B}?\mathbf{S}[\psi(t)]) \tag{33}$$

where ψ is considered to be (ii)-differentiable.

Proof. For arbitrary fixed $r\in[0,1]$ we have

$$\begin{aligned}&\tfrac{1}{B}?\mathbf{S}[\psi(t)]?(\tfrac{1}{B}?\psi(0))\\&=(\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)]-\tfrac{1}{B}\mathbf{S}[\underline{\psi}(0,r)],\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)]-\tfrac{1}{B}\mathbf{S}[\overline{\psi}(0,r)])\end{aligned} \tag{34}$$

We have the following relations

$$\begin{aligned}\mathbf{S}[\overline{\psi}'(t,r)]&=\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)]-\tfrac{1}{B}[\overline{\psi}(0,r)]\\\mathbf{S}[\underline{\psi}'(t,r)]&=\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)]-\tfrac{1}{B}[\underline{\psi}(0,r)]\end{aligned} \tag{35}$$

Hence, we obtain

$$\frac{1}{B}?\mathbf{S}[\psi(t)]?(\frac{1}{B}?\psi(0))=(\mathbf{S}[\underline{\psi}'(t,r)],\mathbf{S}[\overline{\psi}'(t,r)]) \tag{36}$$

If is considered to be (i)-differentiable, so

$$\frac{1}{B}?\mathbf{S}[\psi(t)]?(\frac{1}{B}?\psi(0))=\mathbf{S}[\psi'(t)] \tag{37}$$

Let is (ii)-differentiable. For arbitrary fixed $\alpha\in[0,1]$ we obtain

$$\begin{aligned}&\tfrac{-1}{B}?[\psi(0)]?(\tfrac{-1}{B}?\mathbf{S}[\psi(t)])\\&=(\tfrac{-1}{B}\overline{\psi}(0,r)+\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)],\tfrac{-1}{B}\underline{\psi}(0,r)+\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)])\end{aligned} \tag{38}$$

The above equation can be written as the following relation

$$\begin{aligned}&\tfrac{-1}{B}?[\psi(0)]?(\tfrac{-1}{B}?\mathbf{S}[\psi(t)])\\&=(\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)]-\tfrac{1}{B}\overline{\psi}(0,r),\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)]-\tfrac{1}{B}\underline{\psi}(0,r))\end{aligned} \tag{39}$$

We obtain

$$\begin{aligned}\mathbf{S}[\overline{\psi}'(t,r)]&=\tfrac{1}{B}\mathbf{S}[\overline{\psi}(t,r)]-\tfrac{1}{B}\overline{\psi}(0,r)\\\mathbf{S}[\underline{\psi}'(t,r)]&=\tfrac{1}{B}\mathbf{S}[\underline{\psi}(t,r)]-\tfrac{1}{B}\underline{\psi}(0,r)\end{aligned} \tag{40}$$

So, we have

$$(\frac{-1}{B}\psi(0))?(\frac{-1}{B}?\mathbf{S}[\psi(t)])=(\mathbf{S}[\overline{\psi}'(t,r)],\mathbf{S}[\underline{\psi}'(t,r)]) \tag{41}$$

Hence

$$(\frac{-1}{B}\psi(0))?(\frac{-1}{B}?\mathbf{S}[\psi(t)])=\mathbf{S}([\overline{\psi}'(t,r)],[\underline{\psi}'(t,r)]) \tag{42}$$

Since ψ is (ii)-differentiable, therefore,

$$(\frac{-1}{B}\psi(0))?(\frac{-1}{B}?\mathbf{S}[\psi(t)])=\mathbf{S}[\psi'(t)] \tag{43}$$

Theorem 3: Taking into consideration that Sumudu transform is a linear transformation, so if $\psi(t)$ and $\vartheta(t)$ be continuous Z-number valued functions, moreover k_1 as well as k_2 be constant, therefore the following relation can be obtained

$$\mathbf{S}[(k_1?\psi(t))\oplus(k_2?\vartheta(t))]=(k_1?\mathbf{S}[\psi(t)])\oplus(k_2?\mathbf{S}[\vartheta(t)]) \tag{44}$$

Proof. We have

$$\begin{aligned}&\mathbf{S}[(k_1?\psi(t))\oplus(k_2?\vartheta(t))]\\&=\textstyle\int_0^\infty(k_1?\phi(Bt)\oplus k_2?\vartheta(Bt))?e^{-t}dt\\&=\textstyle\int_0^\infty k_1?\psi(Bt)?e^{-t}dt\oplus\int_0^\infty k_2?\vartheta(Bt)?e^{-t}dt\\&=k_1?(\textstyle\int_0^\infty\psi(Bt)?e^{-t}dt)\oplus k_2?(\int_0^\infty\vartheta(Bt)?e^{-t}dt)\\&=k_1?\mathbf{S}[\psi(t)]\oplus k_2?\mathbf{S}[\vartheta(t)]\end{aligned} \tag{45}$$

Therefore, we conclude

$$\mathbf{S}[(k_1?\psi(t))\oplus(k_2?\vartheta(t))]=(k_1?\mathbf{S}[\psi(t)])\oplus(k_2?\mathbf{S}[\vartheta(t)]) \tag{46}$$

Lemma 1: Assume that $\psi(t)$ is a continuous Z-number value function on $[0,\infty)$, also $\gamma\geq 0$, thus

$$\mathbf{S}[\gamma?\psi(t)]=\gamma?\mathbf{S}[\psi(t)] \tag{47}$$

Proof. Fuzzy Sumudu transform $\gamma?\psi(t)$ is defined as

$$\mathbf{S}[\gamma?\psi(t)]=\textstyle\int_0^\infty\gamma?\psi(Bt)?e^{-t}dt \tag{48}$$

furthermore, we have

$$\textstyle\int_0^\infty\gamma?\psi(Bt)?e^{-t}dt=\gamma?\int_0^\infty\psi(Bt)?e^{-t}dt \tag{49}$$

therefore,

$$\mathbf{S}[\gamma?\psi(t)]=\gamma?\mathbf{S}[\psi(t)] \tag{50}$$

Lemma 2: Assume that $\psi(t)$ is a continuous Z-number valued function, and $\vartheta(t)\geq 0$. Furthermore, if we suppose that $(\psi(t)?\vartheta(t))?e^{-t}$ is improper Z-number Reiman integrable on $[0,\infty)$, then

$$\begin{aligned}&\textstyle\int_0^\infty(\psi(Bt)?\vartheta(Bt))?e^{-t}dt\\&=(\textstyle\int_0^\infty\vartheta(Bt)\underline{\psi}(Bt,r)e^{-t}dt,\int_0^\infty\vartheta(Bt)\overline{\psi}(Bt,r)e^{-t}dt)\end{aligned} \tag{51}$$

Theorem 4: Suppose $\psi(t)$ is a continuous Z-number valued function, also $\mathbf{S}[\psi(t)]=D(B)$, therefore,

$$\mathbf{S}[e^{a_1t}?\psi(t)]=\frac{1}{1-a_1B}D(\frac{B}{1-a_1B}) \tag{52}$$

where e^{a_1t} is considered to be a real value function, also $1-a_1B>0$.

Proof. We have the following relation

$$\begin{aligned}&\mathbf{S}[e^{a_1 t}\,?\,\psi(t)]=\int_0^{\infty}e^{a_1 Bt}e^{-t}\psi(Bt)dt\\&=(\int_0^{\infty}e^{-(1-a_1B)t}\underline{\psi}(Bt,r)dt,\int_0^{\infty}e^{-(1-a_1B)t}\overline{\psi}(Bt,r)dt)\end{aligned}\tag{53}$$

Let us consider $z=1-a_1Bt$, then

$$\begin{aligned}&\mathbf{S}[e^{a_1 t}\,?\,\psi(t)]\\&=\tfrac{1}{1-a_1B}(\int_0^{\infty}\underline{\psi}(\tfrac{Bz}{1-a_1B},r)e^{-z}dz,\int_0^{\infty}\overline{\psi}(\tfrac{Bz}{1-a_1B},r)e^{-z}dz)\\&=\{\tfrac{1}{1-a_1B}\underline{D}(\tfrac{B}{1-a_1B}),\tfrac{1}{1-a_1B}\overline{D}(\tfrac{B}{1-a_1B})\}=\tfrac{1}{1-a_1B}D(\tfrac{B}{1-a_1B})\end{aligned}\tag{54}$$

4. Solving fuzzy initial value problem with fuzzy Sumudu transform method

Consider the following fuzzy initial value problem based on Z-numbers

$$\begin{cases}\phi'(t)=\psi(t,\phi(t)),\\\phi(0)=(\underline{\phi}(0,r),\overline{\phi}(0,r)),\qquad 0<r\le 1\end{cases}\tag{55}$$

where $\psi(t,\phi(t))$ is a Z-number function. By utilizing FST method for Z-numbers, we obtain

$$\mathbf{S}[\phi'(t)]=\mathbf{S}[\psi(t,\phi(t))]\tag{56}$$

The solving process of Eq. (56) is based on the following cases.

Case 1: Assume that $\phi'(t)$ is (i)-differentiable. Base on the Theorem 3 we extract

$$\phi'(t)=(\underline{\phi}'(t,r),\overline{\phi}'(t,r))\tag{57}$$

$$\mathbf{S}[\varphi'(t)]=(\frac{1}{B}?\mathbf{S}[\varphi(t)])?\frac{1}{B}\varphi(0)\tag{58}$$

Eq. (58) can be displayed as the following relation

$$\begin{cases}\mathbf{S}[\underline{\psi}(t,\phi(t),r)]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,\alpha)\\\mathbf{S}[\overline{\psi}(t,\phi(t),r)]=\frac{1}{B}\mathbf{S}[\overline{\phi}(t,r)]-\frac{1}{B}\overline{\phi}(0,r)\end{cases}\tag{59}$$

where

$$\begin{cases}\underline{\psi}(t,\phi(t),r)=\min\{\psi(t,B)\mid B\in(\underline{\phi}(t,r),\overline{\phi}(t,r))\}\\\overline{\psi}(t,\phi(t),r)=\max\{\psi(t,B)\mid B\in(\underline{\phi}(t,r),\overline{\phi}(t,r))\}\end{cases}\tag{60}$$

Accordingly, Eq. (60) can be resolved on the basis of the following assumptions

$$\mathbf{S}[\underline{\phi}(t,r)]=U_1(B,r)\tag{61}$$

$$\mathbf{S}[\overline{\phi}(t,r)]=U_2(B,r)\tag{62}$$

where $U_1(B,r)$, as well as $U_2(B,r)$ are the Z-number solutions of the Eq. (60). By applying inverse Sumudu transform, $\underline{\phi}(t,r)$ as well as $\overline{\phi}(t,r)$ are computed as

$$\underline{\phi}(t,r)=\mathbf{S}^{-1}[U_1(B,r)]\tag{63}$$

$$\overline{\phi}(t,r)=\mathbf{S}^{-1}[U_2(B,r)]\tag{64}$$

Case 2: Assume that $\phi'(t)$ is (ii)-differentiable. Based on the Theorem 3 we extract

$$\phi'(t)=(\overline{\phi}'(t,r),\underline{\phi}'(t,r))\tag{65}$$

$$\mathbf{S}[\varphi'(t)]=(\frac{-1}{B}?\varphi(0))?(\frac{-1}{B}?\mathbf{S}[\varphi(t)])\tag{66}$$

Eq. (66) can be displayed as the following relation

$$\begin{cases}\mathbf{S}[\underline{\psi}(t,\phi(t),r)]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r)\\\mathbf{S}[\overline{\psi}(t,\phi(t),r)]=\frac{1}{B}\mathbf{S}[\overline{\phi}(t,r)]-\frac{1}{B}\overline{\phi}(0,r)\end{cases}\tag{67}$$

where

$$\begin{cases}\underline{\psi}(t,\phi(t),r)=\min\{\psi(t,B)\mid B\in(\underline{\phi}(t,r),\overline{\phi}(t,r))\}\\\overline{\psi}(t,\phi(t),r)=\max\{\psi(t,B)\mid B\in(\underline{\phi}(t,r),\overline{\phi}(t,r))\}\end{cases}\tag{68}$$

Accordingly, Eq. (68) can be resolved on the basis of the following assumptions

$$\begin{aligned}&\mathbf{S}(\underline{\phi}(t,r)=V_1(B,r)\\&\mathbf{S}(\overline{\phi}(t,r)=V_2(B,r)\end{aligned}\tag{69}$$

where $V_1(B,r)$, and $V_2(B,r)$ are the Z-number solutions of the Eq. (68). By applying inverse Sumudu transform, $\underline{\phi}(t,r)$ as well as $\overline{\phi}(t,r)$ are computed as

$$\begin{aligned}&\underline{\phi}(t,r)=\mathbf{S}^{-1}[V_1(B,r)]\\&\overline{\phi}(t,r)=\mathbf{S}^{-1}[V_2(B,r)]\end{aligned}\tag{70}$$

5. Examples

The following examples have been used to narrate the methodology proposed in this paper.

5.1. Example 1

A tank with a heating system is displayed in Figure 1, where $\widetilde{R}=0.5$, the thermal capacitance is $\widetilde{C}=2$ also the temperature is ψ . The model is formulated as follows[10, 37],

$$\begin{cases}\phi'(t)=-\frac{1}{\widetilde{R}\widetilde{C}}\phi(t),\qquad 0\le t\le T\\\phi(0)=[(\underline{\phi}(0,r),\overline{\phi}(0,r)),p(0.8,0.9,1)]\end{cases}\tag{71}$$

By utilizing the FST method based on Z-number we obtain

$$\mathbf{S}[\phi'(t)]=\mathbf{S}[-\phi(t)]\tag{72}$$

$$S[\phi'(Bt)]\odot e^{-t}dt\tag{73}$$

where $0\le B<K$. By considering case 1 for Z-numbers the following relation is obtained

$$\mathbf{S}[\varphi'(t)]=\frac{1}{B}?(\mathbf{S}[\varphi(t)]?\varphi(0))=\frac{1}{B}\mathbf{S}[\varphi(t)]?\frac{1}{B}\varphi(0) \quad (74)$$

Therefore

$$-\mathbf{S}[\varphi(t)]=\frac{1}{B}\mathbf{S}[\varphi(t)]?\frac{1}{B}\varphi(0) \quad (75)$$

Based on the Eq. (59), we have

$$\begin{cases}-\mathbf{S}[\bar{\phi}(t,r)]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r)\\ -\mathbf{S}[\underline{\phi}(t,r)]=\frac{1}{B}\mathbf{S}[\bar{\phi}(t,r)]-\frac{1}{B}\bar{\phi}(0,r)\end{cases} \quad (76)$$

Therefore, the Z-number solution of Eq. (76) is extracted as $[(\mathbf{S}[\bar{\phi}(t,r)],\mathbf{S}[\underline{\phi}(t,r)]),p(0.8,0.94,1)]$ where

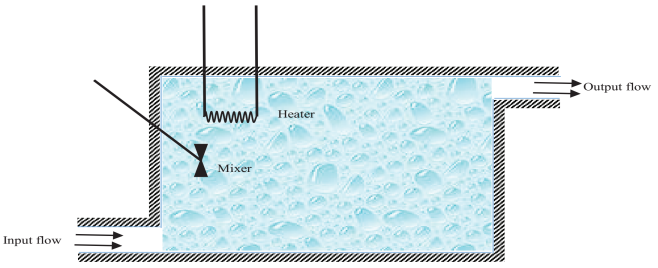


Fig. 1. A tank with a heating system

$$\begin{cases}\mathbf{S}[\bar{\phi}(t,r)]=(\frac{-1}{B^2-1})\bar{\phi}(0,r)+(\frac{B}{B^2-1})\underline{\phi}(0,r)\\ \mathbf{S}[\underline{\phi}(t,r)]=(\frac{-1}{B^2-1})\underline{\phi}(t,r)+(\frac{B}{B^2-1})\bar{\phi}(0,r)\end{cases} \quad (77)$$

By utilizing the inverse Sumudu transform for Z-numbers, we have

$$\begin{cases}\mathbf{S}[\bar{\phi}(t,r)]=\bar{\phi}(0,r)\mathbf{S}^{-1}(\frac{-1}{B^2-1})+\underline{\phi}(0,r)\mathbf{S}^{-1}(\frac{B}{B^2-1})\\ \mathbf{S}[\underline{\phi}(t,r)]=\underline{\phi}(0,r)\mathbf{S}^{-1}(\frac{-1}{B^2-1})+\bar{\phi}(0,r)\mathbf{S}^{-1}(\frac{B}{B^2-1})\end{cases} \quad (78)$$

where

$$\begin{cases}\bar{\phi}(t,r)=e^{t}(\frac{\bar{\phi}(0,r)-\underline{\phi}(0,r)}{2})+e^{-t}(\frac{\bar{\phi}(0,r)+\underline{\phi}(0,r)}{2})\\ \underline{\phi}(t,r)=e^{t}(\frac{\underline{\phi}(0,r)-\bar{\phi}(0,r)}{2})+e^{-t}(\frac{\underline{\phi}(0,r)+\bar{\phi}(0,r)}{2})\end{cases} \quad (79)$$

By considering case 2 for Z-numbers the following relation is obtained

$$\mathbf{S}[\varphi'(t)]=(\frac{-1}{B}\mathbf{S}[\varphi(t)])?(\frac{-1}{B}\varphi(0)) \quad (80)$$

Hence

$$-\mathbf{S}[\varphi(t)]=(\frac{-1}{B}\mathbf{S}[\varphi(t)])?(\frac{-1}{B}\varphi(0)) \quad (81)$$

Based on the above relations, Eq. (71) is illustrated as

$$\begin{cases}-\mathbf{S}[\underline{\phi}(t,r)]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r)\\ -\mathbf{S}[\bar{\phi}(t,r)]=\frac{1}{B}\mathbf{S}[\bar{\phi}(t,r)]-\frac{1}{B}\bar{\phi}(0,r)\end{cases} \quad (82)$$

So, the Z-number solution of Eq. (82) is displayed as $[(\mathbf{S}[\underline{\phi}(t,r)],\mathbf{S}[\bar{\phi}(t,r)]),p(0.8,0.9,1)]$ where

$$\begin{cases}\mathbf{S}[\underline{\phi}(t,r)]=\underline{\phi}(0,r)(\frac{1}{B+1})\\ \mathbf{S}[\bar{\phi}(t,r)]=\bar{\phi}(t,r)(\frac{1}{B+1})\end{cases} \quad (83)$$

By utilizing the inverse Sumudu transform for Z-numbers, we have

$$\begin{cases}\underline{\phi}(t,r)=\underline{\phi}(0,r)\mathbf{S}^{-1}(\frac{1}{B+1})\\ \bar{\phi}(t,r)=\bar{\phi}(0,r)\mathbf{S}^{-1}(\frac{1}{B+1})\end{cases} \quad (84)$$

where

$$\begin{cases}\underline{\phi}(t,r)=e^{-t}\underline{\phi}(0,r)\\ \bar{\phi}(t,r)=e^{-t}\bar{\phi}(0,r)\end{cases} \quad (85)$$

If the initial condition is taken to be a symmetric triangular Z-number as $\phi(0)=[(-a(1-r),a(1-r)),p(0.8,0.9,1)]$, so

Case 1 :

$$\begin{cases}\underline{\phi}(t,r)=e^{t}(-a(1-r))\\ \bar{\phi}(t,r)=e^{t}(a(1-r))\end{cases} \quad (86)$$

Case 2:

Table 1. Approximation errors based on Z-numbers

α	FST	Average Euler
0	[(0.0078,0.0195),p(0.8,0.86,0.94)]	[(0.0138,0.0215),p(0.7,0.8,0.87)]
0.2	[(0.0085,0.0169),p(0.75,0.8,0.9)]	[(0.0188,0.0286),p(0.7,0.8,0.87)]
0.6	[(0.0058,0.0115),p(0.8,0.9,1)]	[(0.0182,0.0198),p(0.7,0.8,0.92)]
0.8	[(0.0091,0.0123),p(0.7,0.75,0.8)]	[(0.0148,0.0189),p(0.6,0.7,0.8)]
1	[(0.0132,0.0132),p(0.7,0.8,0.9)]	[(0.0710,0.0710),p(0.6,0.75,0.87)]

$$\begin{cases}\underline{\phi}(t,r)=e^{-t}(-a(1-r))\\ \bar{\phi}(t,r)=e^{-t}(a(1-r))\end{cases} \quad (87)$$

Approximation errors based on Z-numbers are shown in Table 1. These errors are the differences between the exact and the approximation solutions, for two different methods: FST and Average Euler method [38].

The following formula is utilized to transfer the Z-numbers to fuzzy numbers,

$$\sigma=\frac{\int\phi\pi_P(\phi)d\phi}{\int\pi_P(\phi)d\phi}$$

By taking in to consideration $Z=(B,\tilde{p})=[(0.0078,0.0195),p(0.8,0.86,0.94)]$, we obtain $Z^{\sigma}=[0.0078,0.0195;0.86]$ accordingly $Z'=[\sqrt{0.86}\,0.0078,\sqrt{0.86}\,0.0195]$. Approximation errors based on fuzzy numbers are shown in Table 2.

Figure 2 shows the corresponding error plots based on fuzzy numbers. FST is more accurate than the Average Euler method. Figure 3 and Figure 4 demonstrate the corresponding solution plots based on fuzzy numbers.

By implementing Z-numbers the degree of reliability of the information can be increased [39, 40]. The comparison between the Z-number $Z=[(0.0078,0.0195),p(0.8,0.86,0.94)]$ and fuzzy number $[0.0072,0.0180]$ is displayed in Figure 5. It can be seen that the Z-number incorporates with several information, also the solution related to the Z-number is more precise. The membership function regarding the restriction in the Z-number is considered to be $\mu_{B_Z}=[0.0078,0.0195]$. It can be in probability form.

Table 2. Approximation errors based on fuzzy numbers

α	FST	Average Euler
0	[0.0072,0.0180]	[0.0123,0.0192]
0.2	[0.0076,0.0151]	[0.0168,0.0255]
0.6	[0.0055,0.0109]	[0.0162,0.0177]
0.8	[0.0078,0.0106]	[0.0123,0.0158]
1	[0.0118,0.0118]	[0.0614,0.0614]

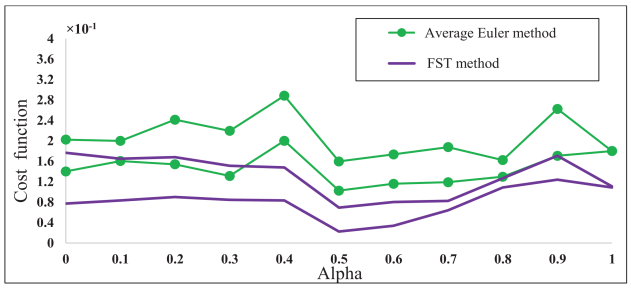


Fig. 2. The lower and upper bounds of absolute errors based on fuzzy numbers

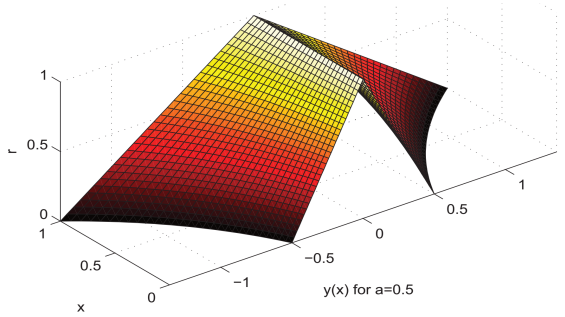


Fig. 3. The solution of FDE under case 1 consideration based on fuzzy numbers

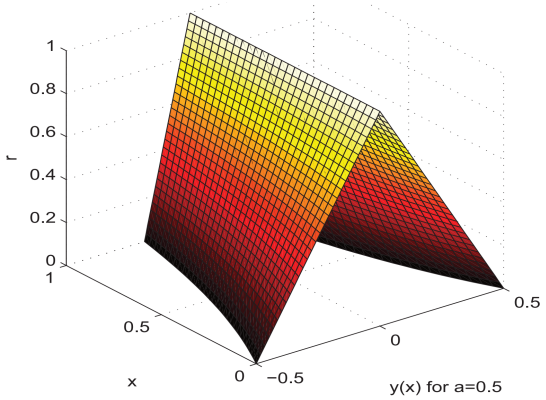


Fig. 4. The solution of FDE under case 2 consideration based on fuzzy numbers

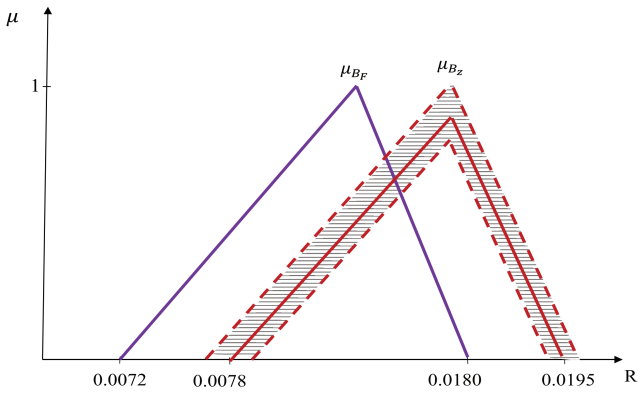


Fig. 5. The comparison between the Z-number and fuzzy number

5.2. Example 2

A tank system is displayed in Figure 6. Suppose $I=t+1$ is inflow disturbances of the tank that generates vibration in liquid level ϕ, also $H=1$ is the flow obstruction, which can be curbed utilizing the valve. $Q=1$ is the cross-section of the tank. The liquid level is illustrated as following relation [41, 42],

$$\begin{cases} \phi'(t)=-\frac{1}{QH}\phi(t)+\frac{I}{Q}, & 0\le t\le T \\ \phi(0)=[(\underline{\phi}(0,r),\overline{\phi}(0,r)),p(0.85,0.91,1)] \end{cases} \quad (88)$$

By utilizing the FST method based on Z-number we obtain

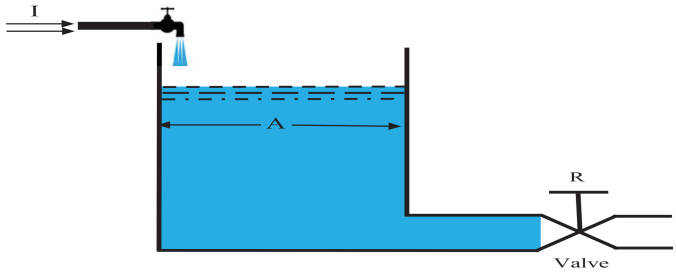


Fig. 6. Liquid tank system

$$-\mathbf{S}[\varphi(t)]=(\frac{1}{B}?\mathbf{S}[\varphi(t)])?(\frac{1}{B}\mathbf{S}[\varphi(0)]) \quad (89)$$

$$\mathbf{S}[\phi'(t)](\int_{a_1}^{\infty}\phi(Bt)e^{'-t}dt \quad (90)$$

By considering case 2 for Z-numbers the following relation is obtained

$$\mathbf{S}[\varphi'(t)]=(\frac{-1}{B}?\mathbf{S}[\varphi(t)])?(\frac{-1}{B}\mathbf{S}[\varphi(0)]) \quad (91)$$

So

$$-\mathbf{S}[\varphi(t)]+\mathbf{S}[t]+\mathbf{S}[1]=(\frac{-1}{B}?\mathbf{S}[\varphi(t)])?(\frac{-1}{B}\mathbf{S}[\varphi(0)]) \quad (92)$$

Based on the Eq. (59), we have

$$\begin{cases} -\mathbf{S}[\underline{\phi}(t,r)]+\mathbf{S}[t]+\mathbf{S}[1]=\frac{1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r) \\ -\mathbf{S}[\overline{\phi}(t,r)]+\mathbf{S}[t]+\mathbf{S}[1]=\frac{1}{B}\mathbf{S}[\overline{\phi}(t,r)]-\frac{1}{B}\overline{\phi}(0,r) \end{cases} \quad (93)$$

Therefore, the Z-number solution of Eq. (93) is extracted as

$$\begin{cases} \mathbf{S}[\underline{\phi}(t,r)]=\mathbf{S}[t]+\mathbf{S}[1]+\frac{-1}{B}\mathbf{S}[\underline{\phi}(t,r)]-\frac{1}{B}\underline{\phi}(0,r) \\ \mathbf{S}[\overline{\phi}(t,r)]=\mathbf{S}[t]+\mathbf{S}[1]+\frac{1}{B}\mathbf{S}[\overline{\phi}(t,r)]-\frac{1}{B}\overline{\phi}(0,r) \end{cases} \quad (94)$$

hence,

$$\begin{cases} \mathbf{S}[\underline{\phi}(t,r)]=(\frac{1}{B+1})\underline{\phi}(0,r)+B \\ \mathbf{S}[\overline{\phi}(t,r)]=(\frac{1}{B+1})\overline{\phi}(t,r)+B \end{cases} \quad (95)$$

By utilizing the inverse Sumudu transform for Z-numbers, we obtain

$$\begin{cases} \underline{\phi}(t,r)=\underline{\phi}(0,r)\mathbf{S}^{-1}(\frac{1}{B+1})+\mathbf{S}^{-1}(B) \\ \overline{\phi}(t,r)=\overline{\phi}(0,r)\mathbf{S}^{-1}(\frac{1}{B+1})+\mathbf{S}^{-1}(B) \end{cases} \quad (96)$$

Where $$\begin{cases} \underline{\phi}(t,r)=e^{-t}\underline{\phi}(0,r)+t \\ \overline{\phi}(t,r)=e^{-t}\overline{\phi}(0,r)+t \end{cases} \quad (97)$$

If the initial condition is taken to be a symmetric triangular Z-number as $\phi(0)=[(-a(1-r),a(1-r)),p(0.85,0.9,1)]$, , so

$$\begin{cases} \underline{\phi}(t,r)=e^{-t}(-a(1-r))+t \\ \overline{\phi}(t,r)=e^{-t}(a(1-r))+t \end{cases} \quad (98)$$

Approximation errors based on Z-numbers are shown in Table 3. These errors are the differences between the exact and the approximation solutions, for two different methods: FST and Max-Min Euler method [38]. Figure 7 shows the Corresponding solution plot based on fuzzy numbers.

Table 3. Approximation errors based on Z-numbers

α	FST	Max-Min Euler
0	[(0.0053,0.0128),p(0.8,0.85,0.9)]	[(0.0108,0.0195),p(0.8,0.9,1)]
0.2	[(0.0019,0.0127),p(0.7,0.81,0.9)]	[(0.0072,0.0138),p(0.6,0.7,0.8)]
0.4	[(0.0031,0.0089),p(0.8,0.9,1)]	[(0.0113,0.0123),p(0.7,0.8,0.91)]
0.6	[(0.0008,0.0027),p(0.8,0.9,1)]	[(0.0043,0.0078),p(0.7,0.8,0.9)]
0.8	[(0.0049,0.0079),p(0.7,0.8,0.86)]	[(0.0069,0.0099),p(0.6,0.7,0.8)]
1	[(0.0109,0.0109),p(0.7,0.8,0.9)]	[(0.0587,0.0587),p(0.6,0.75,0.8)]

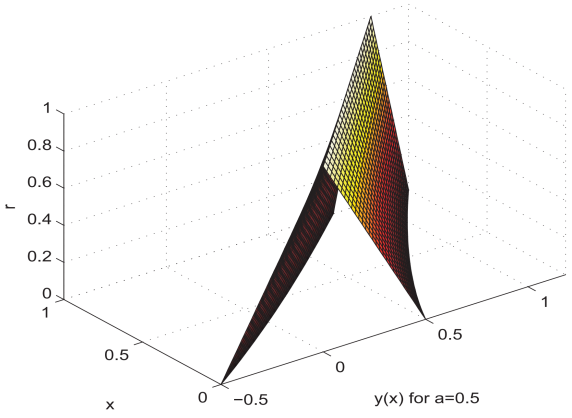


Fig. 7. The solution of FDE under case 2 consideration based on fuzzy numbers

6. Conclusion

In this paper, a novel method based on the FST is proposed in order to find the solution of the first order FDEs on the basis of the Z-numbers. The new method is clarified by utilizing the concept of strongly generalized differentiability. By using the FST method, the FDE converts to an algebraic problem. Some essential theorems are laid down in order to demonstrate the properties of the FST. Two real examples are applied to demonstrate the effectiveness of the proposed technique. This work has a significant contribution in initializing a superior starting point for such extensions.

References

[1] R. Jafari, S. Razvarz, "Solution of Fuzzy Differential Equations using Fuzzy Sumudu Transforms" IEEE International Conference on Innovations in Intelligent Systems and Applications, 84-89, 2017. DOI: 10.1109/INISTA.2017.8001137

[2] D. Zwillinger, "Handbook of differential equations" Gulf Professional Publishing. 1998.

[3] R. Jafari, W. Yu, "Fuzzy Control for Uncertainty Nonlinear Systems with Dual Fuzzy Equations" Journal of Intelligent and Fuzzy Systems. **29**(3), pp.1229-1240, 2015. DOI: 10.3233/IFS-151731

[4] R. Jafari, W. Yu, "Fuzzy Differential Equation for Nonlinear System Modeling with Bernstein Neural Networks" IEEE Access., 2017. doi:10.1109/ACCESS.2017.2647920

[5] R. Jafari, W. Yu, "Uncertainty Nonlinear Systems Modeling with Fuzzy Equations" Mathematical problems in Engineering. **2017**, 2017. doi:10.1155/2017/8594738

[6] S.L. Chang, L.A. Zadeh, "On fuzzy mapping and control" IEEE Trans. Syst. Man Cybern. **2**(1), 30-34, 1972. DOI: 10.1109/TSMC.1972.5408553

[7] D. Dubois, H. Prade, "Towards fuzzy differential calculus: part 3 differentiation" Fuzzy Sets Syst. **8**(3), 225-233, 1982. https://doi.org/10.1016/S0165-0114(82)80001-8

[8] P. Kloeden, "Remarks on Peano-like theorems for fuzzy differential equations" Fuzzy Set Syst. **44**(1), 161-164, 1991. https://doi.org/10.1016/0165-0114(91)90041-N

[9] B. Bede, I.J. Rudas, A.L. Bencsik "First order linear fuzzy differential equations under generalized differentiability" Inform. Sci. **177**(7), 1648-1662, 2007. https://doi.org/10.1016/j.ins.2006.08.021

[10] T. Allahviranloo, M.B. Ahmadi, "Fuzzy Laplace Transform" Soft Computing. **14**, 235-243, 2010.

[11] Z. Ding, M. Ma, A. Kandel, "Existence of the solutions of fuzzy differential equations with parameters" Inform Sci. **99**(3-4), 205-217, 1997. https://doi.org/10.1016/S0020-0255(96)00279-4

[12] V.A. Truong, V.H. Ngo, D.P. Nguyen, "Global existence of solutions for interval-valued integro-differential equations under generalized H-differentiability" Advances in Difference Equations. **1**, 217-233, 2013. https://doi.org/10.1186/1687-1847-2013-217

[13] T. Allahviranloo, N.A. Kiani, M. Barkhor- dari, "Toward the exiistence and uniqueness of solution of second-order fuzzy differential equations" Inform. Sci. **179**, 1207-1215, 2009. https://doi.org/10.1016/j.ins.2008.11.004

[14] F. Hawrra, K.H. Amal, "On fuzzy Laplace transform for fuzzy differential equations of the third order" Journal of Kerbala University, **11**(3), 251-256, 2013.

[15] L. Stefanini, B. Bede, "Generalized Hukuhara differentiability of interval-valued functions and interval differential equations" J. Nonlinear Anal. (TMA). **71**(3-4), 1311-1328, 2009. https://doi.org/10.1016/j.na.2008.12.005

[16] F.B.M. Belgacem, A.A. Karaballi, S.L. Kalla, "Analytical investigations of the Sumudu transform and applications to integral

production equations" Math. Prob. Eng. **103**(3), 103-118, 2003. http://dx.doi.org/10.1155/S1024123X03207018.

[17] Y. Liu, W. Chen, "A new iterational method for ordinary equations using Sumudu transform" Advances in Analysis, **1**(2), 89-94, 2016. DOI: 10.22606/aan.2016.12004

[18] H.M. Srivastava, A. Khalili Golmankhaneh, D. Baleanu, X.J Yang, "Local Fractional Sumudu Transform with Application to IVPs on Cantor Sets" Abstract and Applied Analysis. Article ID 620529, **2014**.1-7, 2014. http://dx.doi.org/10.1155/2014/620529, 2014.

[19] F.B.M. Belgacem, A.A. Karaballi, "Sumudu transform fundamental properties investigations and applications" J. Appl. Math. Stoch. Anal. **2006**, 1-23, 2006. http://dx.doi.org/10.1155/JAMSA/2006/91083

[20] A. Kiliçman, H. Eltayeb, R.P. Agarwal, "On Sumudu Transform and System of Differential Equations" abstr. Appl. Anal, **2010**, 1-11, 2010. doi:10.1155/2010/598702.

[21] N.A. Abdul Rahman, M.Z. Ahmad, "Fuzzy Sumudu transform for solving fuzzy partial differential equations" J. Nonlinear Sci. Appl. **9**, 3226-3239, 2016.

[22] Q.D. Katatbeh, F.B.M. Belgacem, "Applications of the Sumudu transform to fractional differential equations" Nonlinear Studies. **18**(1), 99-112, 2011.

[23] L.A. Zadeh, "Generalized theory of uncertainty (GTU)-principal concepts and ideas" Computational Statistics and Data Analysis, **51**(1), 15-46, 2006. https://doi.org/10.1016/j.csda.2006.04.029

[24] B. Kang, D. Wei, Y. Li, Y. Deng, "A method of converting Z-number to classical fuzzy number" Journal of Information and Computational Science, **9**(3), 703-709, 2012.

[25] L.A. Gardashova, "Application of operational approaches to solving decision making problem using Z-Numbers" Journal of Applied Mathematics, **5**(9), 1323-1334, 2014. DOI: 10.4236/am.2014.59125

[26] R.A. Aliev, A.V. Alizadeh, O.H. Huseynov, "The arithmetic of discrete Z-numbers" Inform. Sci. **290**, 134–155, 2015. https://doi.org/10.1016/j.ins.2014.08.024

[27] B. Kang, D. Wei, Y. Li, Y. Deng, "Decision making using Z-Numbers under uncertain environment" Journal of Computational Information Systems, **8**(7), pp.2807-2814, 2012.

[28] A. Jafarian, R. Jafari, M. Mohamed Al Qurashi, D. Baleanud, "A novel computational approach to approximate fuzzy interpolation polynomials" Springer Plus, **5**, pp.14-28, 2016. https://doi.org/10.1186/s40064-016-3077-5

[29] R. Jafari, W. Yu, X. Li, "Numerical Solution of Fuzzy Equations with Z-numbers using Neural Networks" Intelligent Automation and Soft Computing, 2017. doi: 10.1080/10798587.2017.1327154.

[30] M.L. Puri, D. Ralescu, "Fuzzy random variables" J. Math. Analysis Appl. **114**(2), 409-422, 1986. https://doi.org/10.1016/0022-247X(86)90093-4

[31] H.-C. Wu, "The improper fuzzy Riemann integral and its numerical integration" Inform. Sci. **111**(1-4), 109-137, 1999. https://doi.org/10.1016/S0020-0255(98)00016-4

[32] B. Bede, S.G. Gal, "Generalizations of the differentiability of fuzzy-number-valued functions with applications to fuzzy differential equations" Fuzzy Set Syst. **151**(3), 581-599, 2005. https://doi.org/10.1016/j.fss.2004.08.001

[33] Y. Chalco-Cano, H. Roman-Flores, "On new solutions of fuzzy differential equations" Chaos Solitons Fractals. **38**(1), 112-119, 2006. https://doi.org/10.1016/j.chaos.2006.10.043

[34] L.A. Zadeh, "Toward a generalized theory of uncertainty (GTU) an outline" Inform. Sci. **172** (1-2), 1-40, 2005. https://doi.org/10.1016/j.ins.2005.01.017

[35] R.A. Aliev, W. Pedryczb, V. Kreinovich, O.H. Huseynov, "The general theory of decisions" Inform. Sci. **327**(10), 125-148, 2016. https://doi.org/10.1016/j.ins.2015.07.055

[36] R.A. Aliev, O.H. Huseynov, R.R. Aliyev, A.V. Alizadeh, "The arithmetic of Z-numbers" Theory and applications World Scientific, Singapore, 2015.

[37] R.H. Pletcher, J.C. Tannehill, D. Anderson, "Computational Fluid Mechanics and Heat Transfer", Taylor and Francis, 1997.

[38] S. Tapaswini, S. Chakraverty, "Euler-based new solution method for fuzzy initial value problems" Int. J. Artificial. Intell. Soft. Comput. **4**(1), 58-79, 2014. https://doi.org/10.1504/IJAISC.2014.059288

[39] S. Razvarz, M. Tahmasbi, "Fuzzy equations and Z-numbers for nonlinear systems control", Procedia Computer Science, **120**, 923-930, 2017. https://doi.org/10.1016/j.procs.2017.11.327

[40] R. Jafari, W. Yu, "Uncertain nonlinear system control with fuzzy differential equations and Z-numbers", 18th IEEE International Conference on Industrial Technology, Canada, 890-895, doi:10.1109/ICIT.2017.7915477, 2017.

[41] V.L. Streeter, E.B. Wylie, K.W. Bedford, "Fluid mechanics" New York: McGraw Hill, 1998.

[42] R. Jafari, W. Yu, X. Li, S. Razvarz, "Numerical Solution of Fuzzy Di□erential Equations with Z-numbers Using Bernstein Neural Networks" International Journal of Computational Intelligence Systems, **10**, 1226-1237, 2017. doi:10.2991/ijcis.10.1.81

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 76-81 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Design and Simulation of an RF-MEMS Switch and analysis of its Electromagnetic aspect in realtion to stress

Amna Riaz*,[1], Muhammad Umair Javed Ilam Sindhu[2], Tahir Zaidi[3]

[1]*University of Azad Jammu and Kashmir, Department of Electrical Engineering, Muzaffarabad, 13100, Pakistan*

[2]*Memorial University of Newfoundland, Department of Electrical and Computer Engineering, A1B 3X5, Canada*

[3]*National University of Science and Technology, Department of Electrical Engineering Department, EME College, 13100, Pakistan*

ARTICLE INFO

Article history:
Received: 31 May, 2017
Accepted: 05 September, 2017
Online: 30 January, 2018

Keywords :
RF-MEMS
Residual Stresses
Coplanar waveguide
Skin depth
Return loss
Insertion loss
Isolation
Reliability
Transmission coefficient
Reflection coefficient

ABSTRACT

Microelectromechanical Systems (MEMS) are devices made up of several electrical and mechanical components. They consist of mechanical functions (sensing, thermal, inertial) and electrical functions (switching, decision making) on a single chip made by microfabrication methods. These chips exhibit combined properties of the two functions. The size of system has characteristic dimensions less than 1mm but more than 1μm. The configuration of these components determine the final deliverables of the switch. MEMS can be designed to meet user requirements on any level from microbiological application such as biomedical transducers or tissue engineering, to mechanical systems such as microfluidic diagnoses or chemical fuel cells. The low cost, small mass and minimal power consumption of the MEMS makes it possible to readily integrate to any kind of system in any environment. MEMS are faster, better and cheaper. They offer excellent electrical performances. MEMS working at Radio frequencies are RF MEMS. RF-MEMS switches find huge market in the modern telecommunication networks, biological, automobiles, satellites and defense systems because of their lower power consumptions at relatively higher frequencies and better electrical performances. But the reliability is the major hurdle in the fate of RF MEMS switches. Reliability mainly arises due to the presence of residual stresses, charging current, fatigue and creep and contact degradation. The presence of residual stresses in switches the S-Parameters of the switches are affected badly and the residual stress affects the final planarity of the fabricated structure. Design and simulation of an RF-MEMS switch is proposed considering the residual stresses in both on and off state. The operating frequency band is being optimized and the best possible feasible fabrication technique for the proposed switch design is being analyzed. S-Parameters are calculated and a comparison for the switches with stress and without stress is performed in FEM based HFSS software.

1. Introduction

MEMS technologies are faster, better, and cheaper. They have reduced size and weight, means fewer power consumption and there are lower component counts. MEMS promises superior electrical performances and are built with low cost [1]. They can be produced at massive level. They have advanced functionality and superior reliability. They have new functionality and better system capability.

The ultimate goals of MEMS are to reduce size, cost, weight, power consumptions by increasing the performance of existing macroscopic devices [2]. To make devices never before possible at macroscopic scale. Easy to integrate into systems or modify. They have Small thermal constant and they can be highly resistant to vibration, shock and radiation. MEMS are Microelectro-mechanical system in which the device and structures are fabricated at micro levels using the micro level fabrication process [3]. It is basically an integrated micro devices or systems

*Corresponding Author: Amna Riaz, Email: amna_riaz99@yahoo.com

combining electrical and mechanical components fabricated using Integrated Circuit (IC) compatible processing technique and it range from micron to millimeter [4].

MEMS structures can be a simple system with all static elements or large complicated system with some parts moving and some static elements. Chief concept of MEMS is that there is at least one moving element [5]. MEMS most important elements are micro sensors and micro actuators. Basically these are the transducers that convert electrical to mechanical and vice versa. Micro actuators and micro sensors are widely used in the integrated circuits industry due to low per device cost and other benefits.

2. Radio Frequency MEMS

2.1. Maintaining the Integrity of the Specifications

MEMS can be classified into different types depending upon their mean of actuation. These include acoustic MEMS, optical MEMS and Radio Frequency MEMS (RF-MEMS) [6]. RF MEMS operate in the electromagnetic wave frequencies in the 3 kHz - 300 GHz range [7-8]. This frequency provides maximized efficiency in the circuitry and ability to flow through paths containing insulating material, such as the dielectric insulators of capacitors.

3. Microfabrication of RF-MEMS Switch and Thermal Induction of Residual Stress

In Trento Italy; at ITC-IRST research center; using gold electroplating a process was carried out on micro bridge and RF MEMS; which resulted in obtaining suspended microstructure by photoresist of 3micrometer thickness [2].

3.1. RF Switch Surface (RFS) Micromachining

The RF-MEMS STS whose design is considered in this study is realized through the gold electrodeposition microfabrication

1. At 975C in a wet atmosphere silicon wafer came in contact with thermal oxide of 1000nm thickness and then with nitrogen. Resistors and lines are formed by polysilicon while pattern is formed by dry ditching
2. At 718C silicon oxide is placed with wafer and after that applied to photolithography.
3. A metal is spitted for signal line generation. The general temperature is 400C but when LPCVD oxide is deposited temperature rises by 30C. The area which was free from photoresist; for electrical contact oxide is removed by dry etching then wafer is washed.
4. Wafer is now wet etched and floating metal (Cr/Au) layer is evaporated. 3micrometer thickness of Cr is placed between oxide layer and Au.
5. At 52C electroplating of gold on seed layer is done.
6. At 52C; gold layer is used for strengthen the structure. It's left for 30 minutes a 190C. Finally at 200C structure comes in contact with plasma oxygen for removing extra layers.

3.2. Mechanisms contributing to Residual Stress Induction

For residual stress following steps are made: -

1. Stress In this process; Cr-Au PVD is deposited on photoresist. Seed layer is used for ensuring adhesive property of Au. Although due to thermal instability chromium oxide is formed along with residual stress.
2. Gradient happens due to difference in thermal coefficient of gold and seed layer. Seed layer expands more rapidly and results in tensile residual stress. However micro beams get deformed due to release of Au.
3. Oxygen Plasma ashing helps in avoiding sticking but temperature changes too rapidly forming residual stress. Temperature is maintained at 200C so that Au reached even the narrow paths.

4. Design of RF-MEMS Symmetric Toggle Switches

4.1. On State without stress

The simply designed switch in HFSS is the extension of work presented in "Simulating the Electromagnetic Effects of Thermally Induced Residual Stresses in RF-MEMS Switches". This RF-MEMS switch is designed in FEM based HFSS software. In this switch, (Figure 1) simply a suspended type bridge was designed over the coplanar waveguide transmission line, the design of the switch is as follow. The gold material was assigned to the bridge, this switch was designed just to ensure the results validity with the already present standards in the literature.

Similarly, when the stress was induced in the switch, (Figure 2) the central bridge of the switch deflected upward, which ultimately affected the overall electromagnetic performance of the switch.

4.2. Identify the Headings Design of a simple RF-MEMS Switch in off-state without stress

When the switch is in the off state, the bridge deflects downwards at the center just over the dielectric layer of signal CPW (Figure 3). The side conductors of the CPW are grounded while the signal pass through the central conductor of CPW

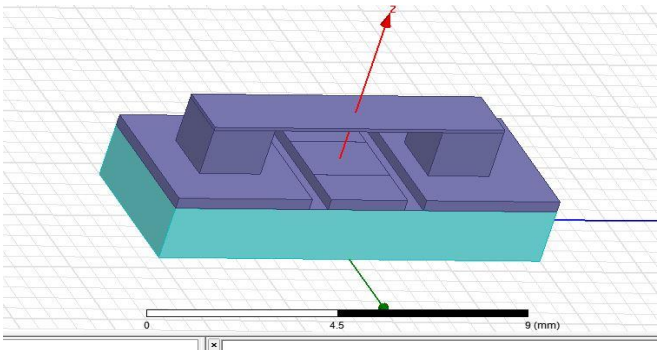


Figure 1: Design of a simple RF-MEMS Switch without stress and excitation

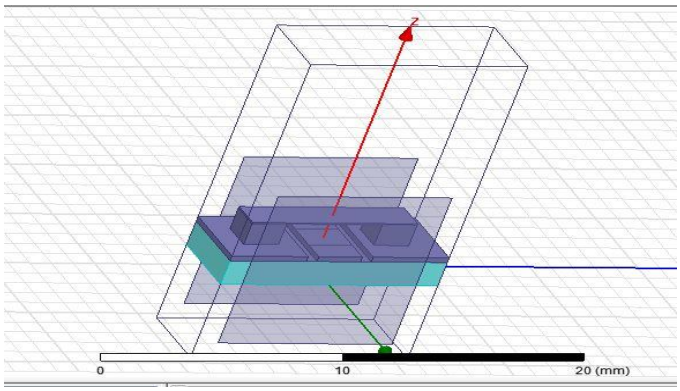


Figure 2: Design of a simple RF-MEMS Switch without stress and excitation and air box

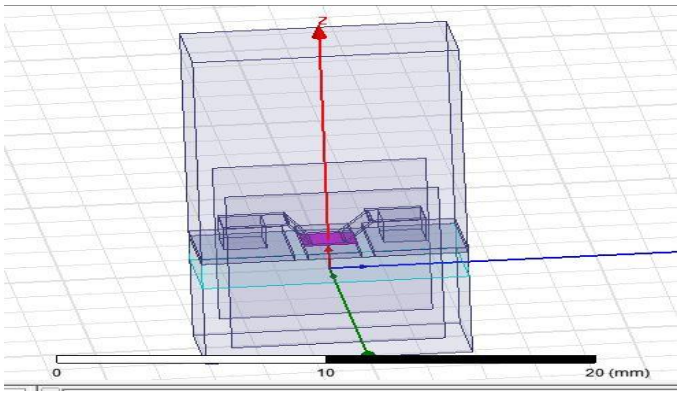


Figure 3: Design of a simple RF-MEMS Switch in off-state without stress

The excitation of the switch is applied through waveport excitations and it is shown in the figure as red shaded side conducting sheet. The waveport connects the central conductor of the CPW with the side conductors, and are used to supply the signal through the switch (Figure 4).

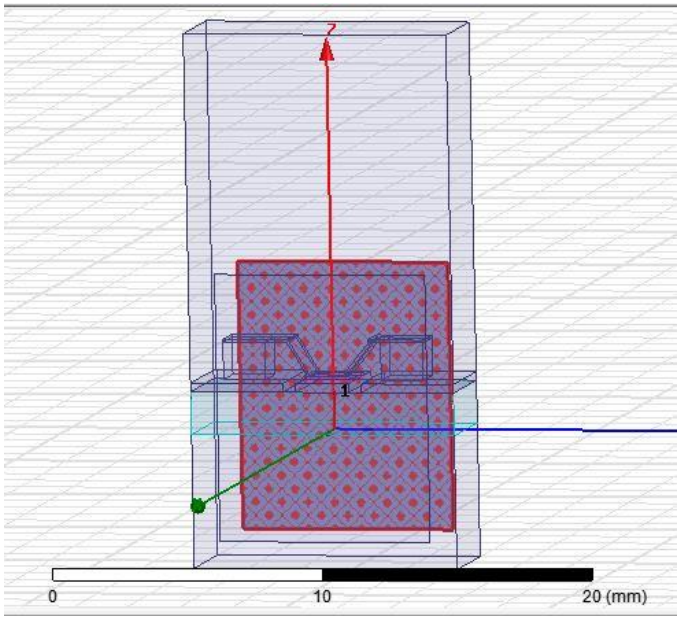


Figure 4: Design of a simple RF-MEMS Switch in off-state without stress with the excitation shown

5. Comparison of simply designed switch with the results already present in literature

5.1. Return Loss of On State

Per the Rangra at Kamal and Jacobi [9], the reflection coefficient of the switch during the on state must be always less than -10 dB, which ensures better switch performances. When we designed a simple switch RF-MEMS switch in HFSS software, the result of reflection coefficient was ascending curve along the X-axis. The result of this switch (Figure 5) is exactly in accordance with existing standards of literature [10, 11] which ensures that our approach was accurate.

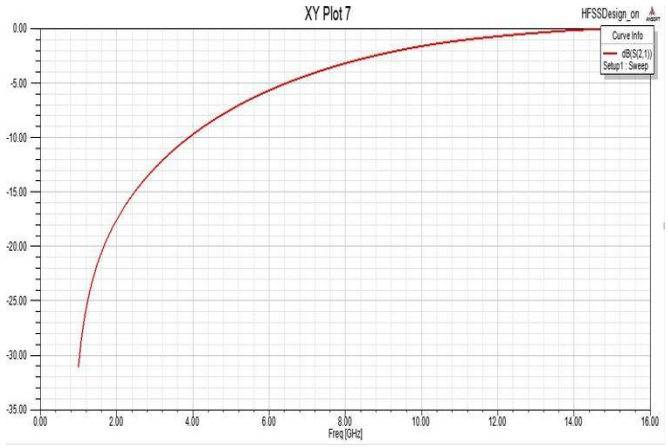


Figure 5: Simulated results of on state return loss

As shown in the graph above, at lower frequencies like between 1.5 GHz to 4 GHz, the reflection was good. Later as the operating frequency range increased, the switch moved to lesser matching or the reflection of the power from the switch increased. From the frequency range of 8GHz to 16 GHz, the switch was not in ideal conditions i.e. the refection increased at higher frequencies. The result already stated in literature (Figure 6) also showed the same trend as shown below [10, 12]. The reflection increased as we move on to higher frequencies.

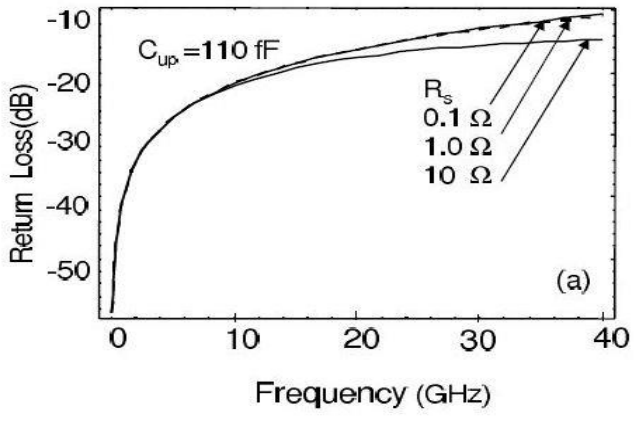


Figure 6: Literature results of on state return loss [2, 10]

5.2. Insertion loss of on state

Ideally, the switch during the on state should have minimum insertion loss, i.e. the value of transmission coefficient should be zero or close to zero. The results which we obtained for a simple RF-MEMS switch in the on state are shown in the graph below, here as the frequency increases, the value of transmission coefficient increases, which predicts lower or poor insertion loss (Figure 7). Ideally the transmission coefficient should be zero as state hypothetically in the literature (Figure 8), but in actual it is not possible to have perfectly zero transmission coefficient [2, 10]. But as the frequency increases, overall there is a decline in the value of transmission coefficient. Which means that at higher frequencies the transmission reduces [13].

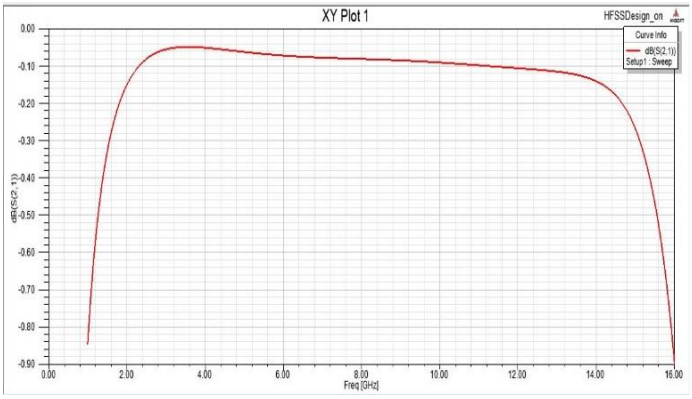


Figure 7: Simulated results of on state Insertion loss

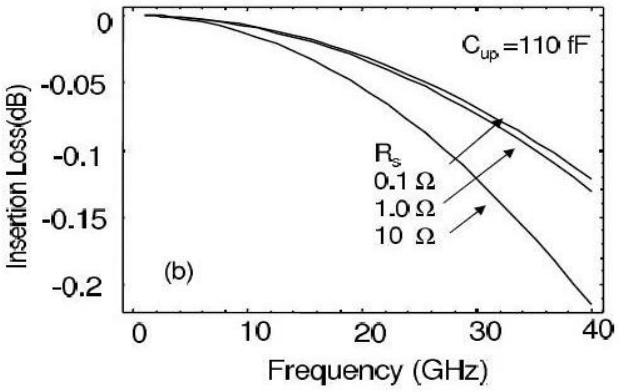


Figure 8: Literature results of on state Insertion loss [2, 10]

5.3. Return Loss of Off-State

Now when the switch is in off-state, the value of reflection coefficient should be greater than -10 dB because at the off-state, switch is not in the working conditions. Maximum power which is incident to the switch gets reflected back [14]. So as the literature predicts, if the operating frequency is taken on X-axis and the transmission coefficients along Y-axis, then the graph should be an ascending curve. So our results of a simple RF-MEMS switch (Figure 9) were found to be exactly similar to those, which were mentioned in the literature (Figure 10). The slight difference in the curve can be explained in terms of the operating frequency ranges i.e. in our graph the limit was between 0 GHz to 16 GHz, whereas in the standard the limit was between 0 GHz to 40 GHz.

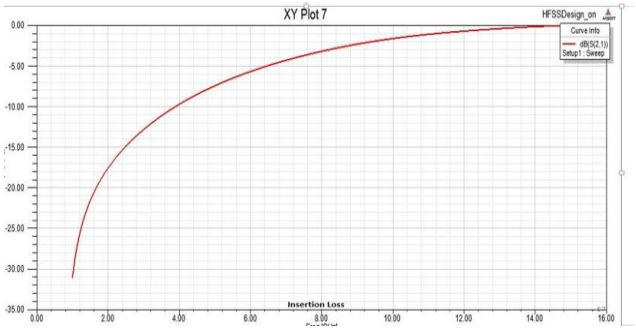


Figure 9: Simulated results of off state return Lost

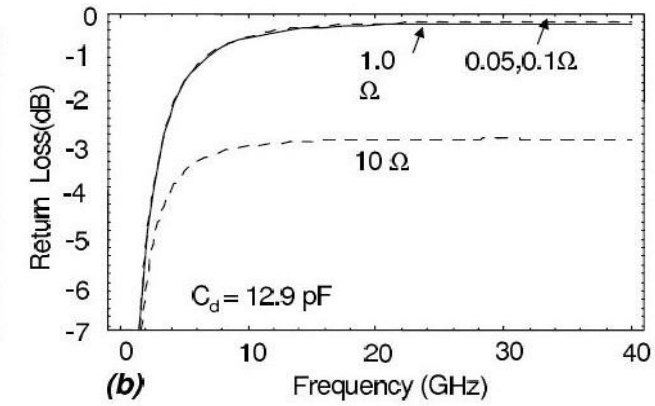


Figure 10: Literature results of off state return Loss. [2, 10]

5.4. Result of Isolation in Off State

According to the set standards, the isolation should be maximum in the off-state. Due to higher isolation, the leakage of power during off-state is minimum [15]. The S_{21} represents the isolation in the off-state. So, here the curve according to the set standard is descending curve, followed by a dip at nearly 15 GHz, and then a gradually ascending curve, after 15 GHz. So, our results of the simple RF-MEMS switch (Figure 11) are in accordance with the pre-defined standards (Figure 12). So, maximum isolation was observed at off-state nearly at 15 GHz.

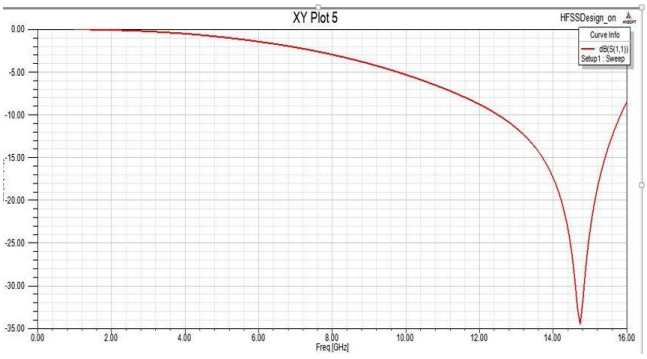


Figure 11: Simulated results of off state Isolation

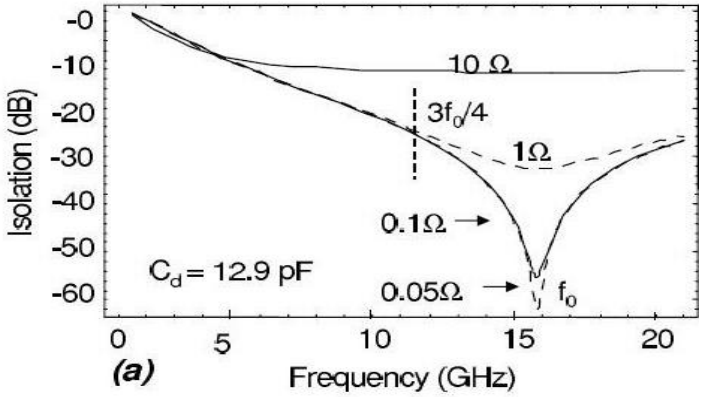


Figure 12: Literature results of off state Isolation. [2, 10]

6. ANALYSIS OF RESULTS

6.1. Without stress

The value of return loss S_{11} in the RF-MEMS switch without stress, shows excellent results. The value of reflection coefficients is much less than -10 dB, which indicates the best matching in the sense of transmitting power. The value of S_{11} lies between -58 dB to -42 dB, which indicates the very less power is reflected. As the value is much lesser than -10 dB, so we can say that switch is having excellent matching or very good reflection loss [16, 17]. Ideally any operating switch must have, reflection loss less than -10 dB, then it is assumed that the electrical performance of the switch is better. If any switch does not have this predefined reflection coefficient, then it means that maximum of the power is wasted or reflected back rather than transmitting through the switch [18]. So, the designed switch in HFSS software, which was already fabricated showed excellent electrical performance in the on state. Their electromagnetic performances are good for the frequency ranges between 2GHz to 16 GHz but, are best between the frequencies of 1.5 GHz to 8GHz. Similarly, the value of transmission coefficient is near 0dB, which means that overall transmission through the switch is excellent. The logic behind stating this is that as we know 0dB is actually equal to 1. It means when the value of transmission is or close to 0dB then the transmission through the switch will be excellent. Again, the operating frequency ranges between 1GHz to 8GHz. Because as the frequency increases, the transmission decreases.

6.2. With Stress

Due to the generation of residual stresses in the RF-MEMS switch, the reflection of the switch is badly affected. The reason is that due to generation of stresses, the bridge deflects upward at the center. Now the air gap between the top central bridge and the ground dielectric increases [19, 20, 21]. Due to rise in air gap the up-state capacitance of the switch is decreased now. As we already explained in the design aspect of RF-MEMS switch that S-parameters are dependent upon the up and down state capacitances. Reflection is dependent upon the up-state capacitances while the isolation is affected by down state capacitances [22, 23, 24]. The results of return loss with stress are shown in Figure 13.

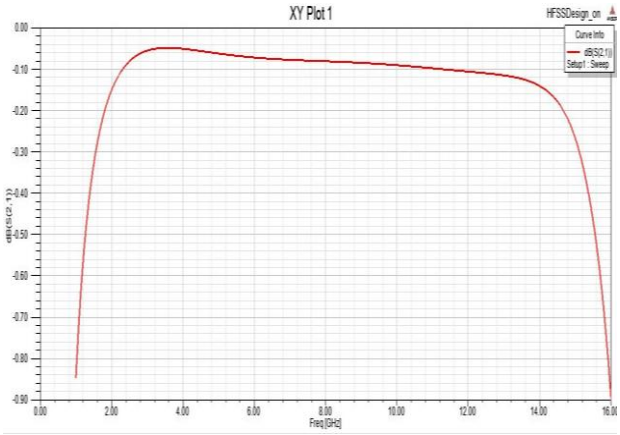


Figure 13: Simulated results of return loss with stress

Similarly, the results for insertion loss are shown in Figure 14.

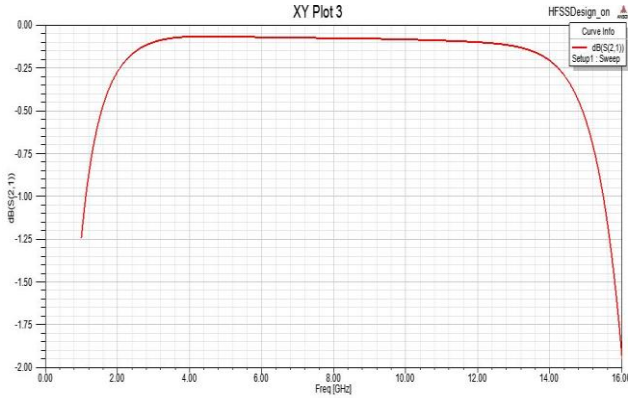


Figure 14: Simulated results of insertion loss with stress

7. Conclusion

So we can conclude that switch was operating efficiently when there were no residual stresses but due to the generation of residual stresses, not only the final planetary of the switch was affected but also the electrical performance of the switch was affected badly. Future work can be done on how to reduce the residual stresses in RF-MEMS switch during microfabrication when the temperature is changed abruptly. Some research can also be produced in the domain of improving reliability of RF-MEMS switches by reducing the effects of charging of dielectric, the effects of creep and fatigue along with the contact degradation. So that it may help to improve the overall efficiency of the RF-MEMS switch.

References

[1] G. Rebeiz, K. Entesari, I. Reines, S. J. Park, M. El-tanani, A. Grichener, and A. Brown, "Tuning in to RF MEMS", Microwave Magazine, IEEE, vol. 10, no. 6, pp. 55-72, 2009.

[2] Solazzi F, Margesin B and Faes A, "Novel Design Solutions for High Reliability RF MEMS Switches", 2010

[3] R. Osiander, M. A. G. Darrin, and J. L. Champion, MEMS and Microstructures in Aerospace Applications. New York, NY, USA: Taylor & Francis, 2006.

[4] G. M. Rebeiz, RF MEMS: Theory, Design, and Technology. John Wiley & Sons, Inc., 2003.

[5] N. Barker and G. Rebeiz, "Optimization of distributed MEMS transmission-line phase shifters-U-band and W-band designs," Microwave Theory and Techniques, IEEE Transactions on, vol. 48, no. 11, pp. 1957 - 1966, Nov. 2000.

[6] A. Verger, A. Pothier, C. Guines, A. Crunteanu, P. Blondy, J. Orlianges, J. Dhennin, F. Courtade, and O. Vendier, "Sub-hundred nanosecond reconfiguration capabilities of nanogap RF MEMS switched capacitor," in Microwave Symposium Digest (MTT), 2010 IEEE MTT-S International, May 2010, pp. 1238 -1241.

[7] H. Newman, J. Ebel, D. Judy, and J. Maciel, -Lifetime Measurements on a High-Reliability RF-MEMS Contact Switch," Microwave and Wireless Components Letters, IEEE, Feb 2008.

[8] Mulloni J, Giacomozzi V and Margesin B 2010 Controlling stress and stress gradient during the release process in gold suspended micro-structures Sensors Actuators A 162 93–9.

[10] B. Pillans, J. Kleber, C. Goldsmith, and M. Eberly, "RF power handling of capacitive RF MEMS devices," in Microwave Symposium Digest, 2002 IEEE MTT-S International, 2002.

[11]J. Muldavin, C. Bozler, S. Rabe, P. Wyatt, and C. Keast, "Wafer- Scale Packaged RF Microelectromechanical Switches," Microwave Theory and Techniques, IEEE Transactions on, vol. 56, no. 2, pp.522 -529, 2008.

[12]A. Grichener and G. Rebeiz, "High-Reliability RF-MEMS Switched Capacitors With Digital and Analog Tuning Characteristics," Mi-crowave Theory and Techniques, IEEE Transactions on, vol. 58, no. 10, pp. 2692 - 2701, 2010.

[13] B. Lakshminarayanan, D. Mercier, and G. Rebeiz, "High-Reliability Miniature RF-MEMS Switched Capacitors," Microwave Theory and Techniques, IEEE Transactions on, vol. 56, no. 4, pp. 971 -981, 2008.

[14] D. Mardivirin, A. Pothier, A. Crunteanu, B. Vialle, and P. Blondy, "Charging in Dielectricless Capacitive RF-MEMS Switches," Microwave Theory and Techniques, IEEE Transactions on, vol. 57,no. 1, pp. 231 -236, 2009.

[15] T. Rijks, J. van Beek, P. Steeneken, M. Ulenaers, J. De Coster, and R. Puers, "RF MEMS tunable capacitors with large tuning ratio," in17th IEEE International Conference on. MEMS, 2004.

[16] F. Casini, P. Farinelli, G. Mannocchi, S. DiNardo, B. Margesin, G. De Angelis, R. Marcelli, O. Vendier, and L. Vietzorreck, "High Performance RF-MEMS SP4T switches in CPW technology for space Applications", in Microwave Conference (EuMC), 2010 European, 2010, pp. 89 -92.

[17] I. Reines, S. J. Park, and G. M. Rebeiz, "Compact Low-Loss Tunable Band Bandstop Filter With Miniature RF-MEMS Switches," Microwave Theory and Techniques, IEEE Transactions on, vol. 58, no. 7, pp. 1887 -1895, 2010.

[18]A. Ocera, P. Farinelli, F. Cherubini, P. Mezzanotte, R. Sorrentino, B. Margesin, and F. Giacomozzi, "A MEMS-Recon_gurable Power Divider on High Resistivity Silicon Substrate," in Microwave Symposium, 2007. IEEE/MTT-S International, 2007, pp. 501 -504.

[19] M. Kim, J. Hacker, R. Mihailovich, and J. DeNatale, "A DC-to-40 GHz four-bit RF MEMS true-time delay network," Microwave and Wireless Components Letters, IEEE, vol. 11, no. 2, pp. 56 -58, Feb.2001.

[20] T. Vaha-Heikkila, J. Varis, J. Tuovinen, and G. Rebeiz, "A 20-50 GHz RF MEMS single-stub impedance tuner," Microwave and Wireless Components Letters, IEEE, vol. 15, no. 4, pp. 205 - 207, 2005.

[21] Rebeiz G M 2003 RF MEMS: Theory, Design and Technology (New York: Wiley)

[22] W. Auer, E. Hettlage, and G. Ru_, "RF-switches: Application and design," in ESA, Proceedings of the 3rd European Space Mechanisms and Tribology Symposium p 191 -195 (SEE N88-21191 14-18), Dec.1987, pp. 191-195.

[23] R.C. Hibbeler, "Mechanics of Materials", Edition 9, pp 153

[24] Fang W, Lee C H and Hu H H 1999 On the buckling behavior of micromachined beams J. Micromech. Microeng.9 236–44

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 82-93 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Perfect Ecosystem for Learning? Modern Thoughts for Organizing Higher Education

Pasi Juvonen*, Anu Kurvinen

Business and Culture, Saimaa University of Applied Sciences, 53850, Lappeenranta, Finland

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
experimental development ecosystem, team entrepreneurship, team learning

ABSTRACT

The background of the article lies in South Eastern Finland, in Lappeenranta, where an active University campus has attracted a group of ICT startups as well as SME's in the field to collaboration. The novel education approach also has another remarkable role as a developer and source of innovation. An experimental development ecosystem (EDE), where learning of knowledge, skills and character are combined, is presented in the article. Also future paths of the EDE are discussed. Companies in the ICT field worldwide are in constant need of competent experts who are ready to adopt the new tools and, at the same time, have an entrepreneurial mindset. We argue that inspiring students to learn through appropriate learning methods and providing them with a modern learning environment comes first. ICT tools, applications and systems to support learning objectives come second. The model presented in the article has been studied for some years as action research. The learning methods that have been found beneficial in IT and marketing bachelor education have been spread to other bachelor and master study programs as well. Results from the data show that students who study as team entrepreneurs have learned content knowledge, meta-skills and reflection skills via the learner-centric methods used in EDE. They have also been inspired to employ new ICT tools and applications to support their learning and project work.

1. Introduction

This article is an extension of the work originally presented in Educon 2017 in Athens in April 2017 [1].

The change in digitalization has been rapid and the change speed has been increasing during the recent decade. Since Facebook, we have seen plenty of new platform business models. Only some of them have succeeded and have been able to capture value for themselves. However, those who have succeeded have been able to capture themselves almost the whole market.

At the same time, many conventional businesses have run into severe challenges caused by rapid digitalization. To give an example, The Finnish Posti – a post delivery company owned by the government – has in recent years laid off employees due the diminishing amount of traditional mail deliveries. At the same time, agile delivery companies have challenged the Finnish Posti by offering customers faster delivery and several add-on services. The Finnish Posti has tried to find new business elsewhere, e.g. from lawn moving services to offering basic health care services.

Another example is the Finnish national railway company VR Group, which has held a monopoly for a long time (over 150 years). During the recent years, discussion about free competition in railway traffic has increased, and current plans may be deployed in the 2020s. Already, these plans have caused major reorganizations and layoffs within the VR Group.

There are more such examples found elsewhere in the world, e.g. Eastman/Kodak. When we consider what will happen in higher education, we can foresee what is going to happen in the future. How will platforms such as Khan Academy, Coursera and Udemy, and other more sophisticated platforms developed after them, change the current education systems? How many teachers will be needed in future to teach basic principles of programming, for example, when high-quality content can be freely loaded via learning platforms?

For us as educators and researchers an important question is: Are we going to build a shelter and try to resist the change, or would it be wiser to build a windmill and try to utilize the winds of change as well as possible?

*Corresponding Author: Pasi Juvonen, Email: pasi.juvonen@saimia.fi

In the world of today, there is plenty of information available. Thus, one has to be able to think critically, have skills to synthesize and put the information into action in a wise way. According to Fadel et al. [2], in order to deepen the learning process in the three essential dimensions – knowledge, skills and character qualities – an important dimension is needed – meta-learning. This means that there are some internal processes required for our learning, namely reflection and adaptation of our learning. Figure 1 presents the framework for the 21st century learner and for the curriculum redesign that tries to answer to these needs.

Our practical experiences with the same philosophy started several years ago. We – a group of lecturers - decided to start building a learning environment which would enable utilizing knowledge gained from different sources, combine theory, learning by doing and reflection, and make rapid changes possible when learning needs a change, without rethinking and redesigning the whole curriculum. This learning environment also provides practical measures to show students' progress.

The current curriculum (presented in section 2.1) is a result from coaching altogether 11 student teams comprising altogether more than 150 students from two areas of specialization - information technology and marketing. The development of the curriculum took place in 2009 - 2016. Since its early steps in 2009, this learning environment has expanded into a learning ecosystem that we call Experimental Development Ecosystem (EDE). This article describes the ecosystem and its pedagogical background, how the learning is organized within it, and the operational level practices that enable it. In this article, we present and discuss our current state of the art with the EDE, ongoing development activities as well as future development paths.

The article is organized as follows. Chapter 2 presents the results from a literature study about the requirements in bachelor education in Finland and in the OECD context. It also presents the current state of the art with the Experimental Development Ecosystem. Chapter 3 discusses data collection and data analysis. Chapter 4 lists some observations based on the data, and finally, Chapter 5 summarizes the findings and discusses our probable future paths on the subject.

2. Rapid changes will require rearrangements in organizing learning environments

Discussion on a right balance between studying explicit content that will, depending on the subject, easily become out-of-date, and meta-skills that are useful but, at the same time, may leave learners with an experience of not learning anything specific has been going on for a long time. Skills that are easy to teach and learning that is easily measured involve skills that are easily automated [3]. We are, to a certain extent, educating young people for future professions that do not exist when decisions about students' curricula are made. These new professions emerge (and some others disappear) while students are studying for their diploma.

Learning is less about reproducing content knowledge. It is more about extrapolating what we know in novel situations [3]. In future, more employees with versatile skills are needed and, at the same time, fewer specialists with deep expertise in one subject are needed. Communal learning skills and team working in multidisciplinary working groups or teams are important [3].

The demands presented for undergraduate education are versatile. Based on a study of literature [3-10], an undergraduate student needs at least the following skills:

- team working
- communal learning
- problem solving
- creativity and innovativeness
- critical thinking
- decision making
- leadership and self-leadership
- shared expertise
- reflection on one's values, and social and emotional skills.

At the same time, learning environments built for supporting learning should

- offer versatile methods for learning
- diminish teacher-led methods
- provide coaching for the constantly changing world
- foster entrepreneurship
- enable running pilots (e.g. establishing cooperatives).

[3-10].

To get an overall picture of what should be considered when organizing higher education in 2017, we looked for a framework for identifying knowledge, skills, meta-level skills and methods. We also wanted to make a cross-section of the EDE compared to other frameworks. The framework we chose is presented in Figure 1.

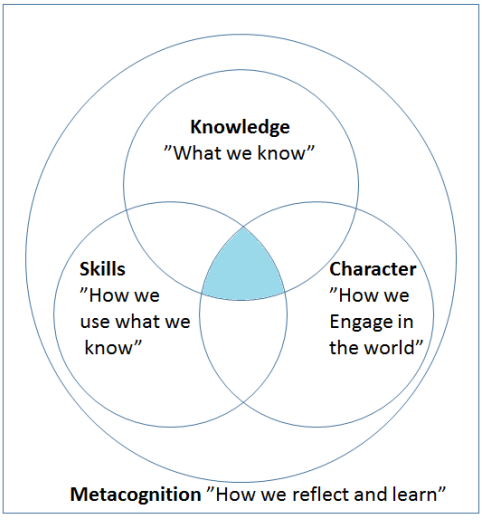


Figure 1. The framework for 21st century education by Centre for Curriculum Redesign [2].

According to [8], there are four forces that lead the learners towards new ways of learning for life in the 21st century. These forces are:

1. Knowledge work – employees who use brain power as well as digital tools for creating new solutions collaboratively in teams

2. Thinking tools – knowledge workers use a set of digital tools, devices and services
3. Digital lifestyles – today's student generations are born into the digital society, and grow up with the digital devices
4. Learning research – the latest research on learning during the last three decades has deepened the understanding of learning processes (See [34 – 37], [39-40], and [51-60] for more)

2.1. State of the art at our UAS – Experimental Development Ecosystem (EDE)

Since 2009 we have been developing a new learning environment, combining studying content knowledge (theory), learning by doing (practice), and employing dialogue [11, 12] in knowledge sharing, knowledge creation and reflection. Together with several local companies and municipalities, we have been able to build the Experimental Development Ecosystem. The current ecosystem offers our undergraduate students an excellent platform for both studying content knowledge, applying this knowledge in real customer projects, and reflecting what has been learned by doing with other team members and the team coach.

Since 2013 we have been developing the EDE in parallel with several RDI projects supported by the Finnish Funding Agency for Innovation (TEKES), Saimaa UAS and LUT University. An "easy-to-start" cooperation between stakeholder groups with little or no thresholds has been a cornerstone for the further development of the EDE.

The EDE was originally created and developed as a social innovation; it is a novel way to organize bachelor education for IT and Business Administration students specializing in marketing at Saimaa UAS. Tiimiakatemia [13] in Jyväskylä and Proakatemia [14] in Tampere were studied as models of how the core structures of team learning can be established. Both of the above academies are specialized in entrepreneurship. An extensive study of how team learning and team entrepreneurship has been organized at Saimaa UAS with IT students to support entrepreneurship education was carried out by Juvonen in 2014 [15].

In 2014 there was no other learning environment designed to support entrepreneurship education in IT Bachelor education [15]. The current employment of the EDE seems to be the only implementation of team entrepreneurship where a student can specialize in marketing (other deployments focus on entrepreneurship overall) and study as a team entrepreneur combining theory, practice and reflections within the EDE.

Studies on learning environments designed for supporting entrepreneurship education in higher education [16- 31], focus on promoting entrepreneurship by fostering either the mindset or skills needed in entrepreneurship, or focus on increasing the status of entrepreneurship as a career choice. Most of the studies we found from Scopus and Science Direct databases (from year 2010 until now) had a narrow focus either on a single course or group of students or a certain technique to foster entrepreneurial skills or thinking. We summarized these studies as follows:

- Applications, tools and methods for fostering entrepreneurial mindset, skills or intentions [19 – 25]
- Evaluation of entrepreneurship education programs overall [26, 27], perceived of value of entrepreneurship education program [27], evaluation of methods used in entrepreneurship education [28]
- How demand and supply meet on entrepreneurship education [30 - 32]
- Student entrepreneurship [33]

Applications with quite a similar approach than the EDE described in this study were found only at Lund University in Sweden and at University of Southern Denmark [18]. At Lund University, practice-based courses and projects are offered, and best practices are spread within an entrepreneurial ecosystem. At the University of Southern Denmark, there are many entrepreneurship education related courses, where different learning methods are used. Most of them are extra-curricular courses, which complicates student participation. Furthermore, basic concepts are taught in a conservative way and practical issues are learned via intensive courses, or via "real entrepreneurs" as visiting lecturers. An active reflection process has been found effective to unleash creative and innovative thinking potential [18].

Students in these two applications are not studying as team entrepreneurs, so they are not running and developing their own enterprise while they are studying. Rather they are participating in a series of teacher-led courses. Based on these differences, the current deployment of the EDE described in this study is considered as a novel approach to entrepreneurship education.

Also concepts of transformational learning [34] have been applied to service-learning while performing service work in [35]. The EDE is more than a team learning environment (Figure 2).

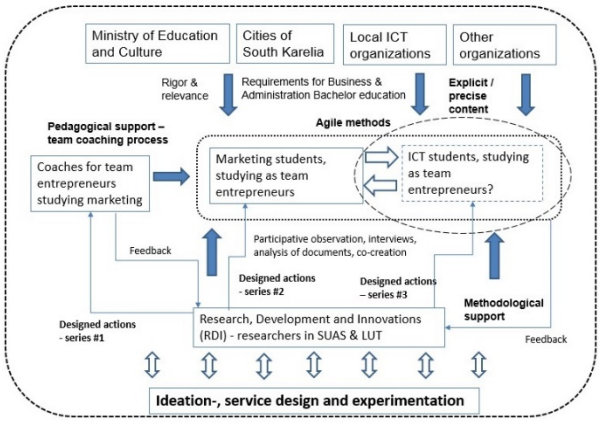


Figure 2. The Experimental Development Ecosystem. Adapted from [36].

The novelty of our approach (the EDE) is in combination of several factors. At the same time this combinations derives the EDE from the approaches found in literature. These factors are as follows:

- Students learn as team entrepreneurs by running a company (a cooperative) they own by themselves
- Focus for students is on specializing in learning marketing in versatile ways
- Experimental learning is emphasized - theory, practice and reflection is involved all the time

- A team coaching process where team development, team performance, and team leadership body of knowledge is applied to support learning
- Several stakeholder groups are involved in the implementation of the EDE
- Continuous cooperation between team entrepreneurs, team coaches, and researchers
- The curriculum has been rebuild to support the team learning and team entrepreneurship and it is further developed based on the experiences gained

Currently, Business Administration students who choose to specialize in marketing study their first year in a conventional way, enrolling on conventional study courses. After choosing the specialization in marketing, they continue their studies for two and half years as team entrepreneurs (Figure 3).

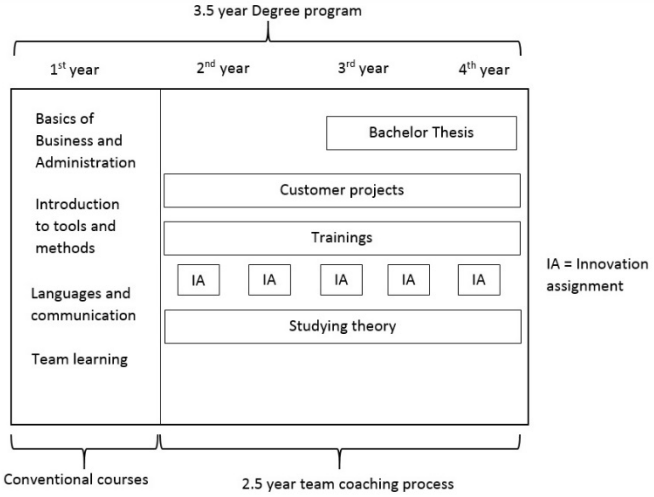


Figure 3. Overview of the curriculum for marketing students studying as team entrepreneurs.

The students who choose to specialize in marketing establish a team enterprise (in the form of a cooperative) and run its operations together during their studies. Their studies consist of studying theory (reading books and articles), carrying out customer projects for real customers for real money, and trainings, where a team coach (personnel of Saimaa UAS) coaches the team members and the team enterprise for team development and performance.

Progress of the team entrepreneurs' studies is measured by five practical measures. Four of these measures show the balance between theory, learning by doing, and sharing knowledge, and reflection. These measures are:

- amount of book points (3 book points equal to 1 ECTS point)
- amount of trainings (in hours, 133 hours equal to 5 ECTS points)
- amount of projects (in hours, 80 hours equal to 3 ECTS points)
- amount of innovation assignments (in hours, 80 hours equal to 3 ECTS points).

To give an example, when a team entrepreneur participates in every training session, which take place twice a week, she will get 5 ECTS points for trainings in a half year. Furthermore, when a team entrepreneur works for 16 hours per week in projects, she gets about 6 ECTS points for projects in a half year. Innovation sessions are held at least twice a year, and participating in them increased the ECTS points. The amount of book points is the only individual measure for the team entrepreneurs. Every team entrepreneur has to complete at least 102 book points, which equals 34 ECTS points. This means that every team entrepreneur has to read at least two business books per month. Book trainings are held twice a month, and those who have read a book and returned a book essay before a book training session are allowed to participate.

The measures described above produce the overall results that are monitored at Saimaa UAS, i.e. the numbers of students who complete at least 55 ECTS points per academic year (1.8 – 31.7). This measure is used by the Finnish Ministry of Education and Culture to monitor study progress in every university of applied sciences in Finland.

The team entrepreneurs are an important element of the ecosystem as collaboration partners and as a scalable source of creativity and innovativeness. Several team entrepreneurs have been recruited already during their studies by local companies. When a company and a team entrepreneur start cooperation on a customer project and continue cooperating on an internship scheme and/or Bachelor's thesis project, it is common, based on our experiences, to continue cooperation after the student has graduated.

The curriculum has been adjusted to make close cooperation with local companies and other organizations possible and fluent. When local companies and other organizations are not able to produce the knowledge they need on their own, they can ask team enterprises for help. A usual method of helping the local organizations is an innovation assignment (IA in Figure 2), where new knowledge is produced within a 12- or 24-hour time limit. Methods of experimental development are used in these assignments in order to create fast, concrete and applicable development ideas based on companies' current needs.

Innovation assignments serve several purposes. For local organizations, they function as a method of rapidly testing their assumptions on a certain topic. For team entrepreneurs, they provide an environment to develop problem-solving skills. Moreover, the assignments function as a measure of team development and substance skills. For all parties, the innovation assignments make it possible to find further cooperation opportunities. In most cases, the organizations participating in an IA will become cooperation partners in the EDE. Sometimes they also take part in RDI projects. Prototypes and/or concepts are usual outcomes of IA's, and prototyping is an important step in an experimental development process. This is the phase that can make a difference in comparison to the traditional workshops where the ideas are easily forgotten after the workshop. It is also easier for the client company to understand the idea and its possibilities for the company after seeing the prototype.

Also research, development, and innovation (RDI) projects support the current EDE model by offering team entrepreneurs

opportunities to work as research assistants. As research assistants, they learn about research methods, gain an understanding of organizing research processes and ways of producing value for local companies [15]. Acting as research assistants in an RDI project has led to employment via internships and/or Bachelors' thesis projects.

2.2. Pedagogics used in the Experimental Development Ecosystem (EDE)

The pedagogics used in the EDE follow ideally the cycle of learning as described by Kolb [37], where learning is described as a process: Firstly, concrete experience is gained, and as the next step in the process, the experience is reflected on. After that comes learning from the experience through abstract conceptualization, and finally testing the newly adopted knowledge and skills through active experimentation.

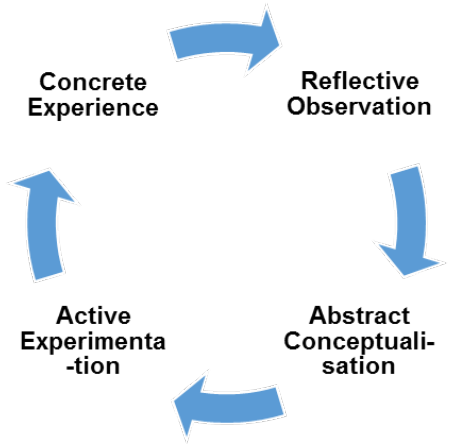


Figure 4. Cycle of learning, adapted from [35].

In constructivism, the learning of a human being is understood as a constant process where individuals are learning or creating their own understanding based on interaction between what they already know and believe, and ideas and knowledge with which they come into contact [38]. Constructivist learning involves at least the following five areas: 1) the educator's attention to the learners, the students and their backgrounds, 2) dialogue facilitation with the group with the purpose of creating a shared understanding of the topic, 3) planned or unplanned introduction of formal theory into the discussion, 4) creating opportunities for the students to challenge or change the existing beliefs and conceptions by using tasks that are structured in a way that makes this possible, and 5) developing students' awareness of their level of understanding and the learning process [39]. In addition to constructivism, team learning, as well as open and honest dialogue, are proven to support the learning objectives.

In the EDE pedagogics, the complexity of companies' operating environments can be learned via concrete experience. At the same time, the theories learned are linked to the reality. Here the principles of building an innovative learning organization are utilized [40]. When considering the current needs of the ICT field, ICT students require more insight into business operations, and practicing business. One way of supporting this is working together in shared projects with team entrepreneur students. Working in multi-disciplinary and heterogeneous project groups is analogous to working life experience, as project teams consist of experts from different fields. Reflection is an elemental part of the learning process and reflective dialogue is used as a pedagogical tool for deepening the learning experience. The researchers working in the ecosystem support the learning process, as they help the students in abstract conceptualization of the phenomena learned. The persons working in the ecosystem need coaching skills that can be applied to the EDE's needs as is required.

These modern thoughts about higher education will also require new skills from those who are employing the system with students. Instead of transferring information to the team entrepreneurs, the emphasis in the team coach's work is rather on helping the process of team development and facilitation of the learning of the information and skills that the team members need for their collaboration and learning (see [41] for more). These teamwork skills and competencies include adaptability, shared situational awareness, performance monitoring and feedback, leadership and team management, interpersonal relations, coordination, communication, and decision making. When these are managed with success, high commitment to learning can be achieved. High commitment usually leads to high performance [42].

A team coach should also make oneself familiar with different styles of consultation. These styles include acceptant, catalytic, confrontational, and prescriptive styles [43]. In the acceptant style, feelings are involved, and this style can be described as emphatic listening. The catalytic style helps the coachee to make decisions. In the confrontational style, the team coach points out what will follow if the coachee continues with her current behavior. The prescriptive style is common in conventional pedagogics. It gives direct advice; however, it does not offer the coachee any opportunity for growth. All these styles are needed in team coaching.

3. Data collection and analysis

The research framework for investigating all the development activities concerning the Experimental Development Ecosystem has been action research [44, 45]. During September 2015 – September 2017 there were several mini-cycles where designed development activities were carried out in different parts of the EDE, and the team entrepreneurs were active participants in many of these development activities.

The outcome of these development activities was monitored through participative observation. The team coaches are active agents for change when they act with student team entrepreneurs. By choosing to use qualitative methods of inquiry, the authors have, at the same time, committed themselves to continuous reflection of their own values and how they affect the research.

The field notes have provided valuable qualitative data, which has been analyzed with other researchers. Two other team coaches have been involved in the sense-making process of how to utilize

the current EDE in ICT education and how to foster cooperation between team entrepreneurs specialized in marketing and ICT bachelor students.

Survey and theme-based interviews during a one-year period between 10/2016 - 9/2017, followed by participative observation, have served as the main data collection methods for the study. An overview of the collected data for the study is presented in Table 1.

Table 1: Overview of the data.

Data collection method	Amount	Timeline	More information
Theme-based interviews	12	10/2016	13 specialists in 12 interviews, from different stakeholder groups involved with the EDE, were interviewed to find out development targets for the EDE.
Participative observation sessions	Over 50	Training sessions, four hours twice a week between 9/2016 – 9/2017	Team coach acts as an active agent for development and provides examples of digital tools for team entrepreneurs.
Theme-based interviews	34	3/2016 – 9/2017, lasted 15 – 40 minutes	Evaluation of the role of the ICT / Digital tools and reflections on why and how team entrepreneurs have been deploying them.
Survey	14 replies	8/2017 – 9/2017	E-mail survey followed by interviews as a part of development discussions.

Methodologically, this article is a partly descriptive and partly explorative case study [46, 47]. It presents the current implementation of the EDE, explores its possible development paths, and finally describes how the ICT Bachelor curriculum could benefit from it.

The basic assumption of the authors is that every research is value-laden and biased. By choosing to use qualitative methods for the inquiry, the authors have, at the same time, committed themselves to continuous reflection on their own values and how they affect the research. In this study, the objective has been the further development of the current EDE and integration of the ICT Bachelor education into it at some level. Therefore, there is an inbuilt bias in the observations and interventions made. However, the authors present the interpretation based on the analysis and the process of conducting research transparently and leave judgement of the validity of study to the reader.

The results of the study have been discussed with three team coaches and two researchers. Luckily, the team coaches share an office at the campus. This has helped to test inner validation of the observations made based on the interviews. The data for the article consists of qualitative interview material (12 specialist interviews with open-ended questions, notes on direct and participative observations, 10+ steering group meetings, researcher workshops, other workshops and meetings), and several unofficial discussions with colleagues and administrative staff at the Saimaa UAS campus and elsewhere where the authors have been actively involved in development activities.

Multiple sources of data and close cooperation between the two authors made it possible to utilize both investigator triangulation and data triangulation [48]. The triangulation of data and researchers has helped to test inner validation of the observations made based on the interviews and participative observation sessions. The data was analyzed applying the principles of grounded theory [49, 50]. The grounded theory analysis includes three main phases: open coding, axial coding, and selective coding [49], and the method requires the researcher theoretical sensitivity [51]. The researcher cannot force the data, but instead she has to let the data "speak". Naturally, this phase is extremely hard in cases where interviewers have a lot of pre-existing knowledge about the subject studied. The process can be made easier by asking the same open-ended questions about the subject studied from all the interviewees and carefully listening to and reporting their expressions.

In the open coding phase, interesting phenomena in the data were highlighted. In this study, the interview notes were first gathered into one text file and then analyzed by the two researchers. In the axial coding phase, the interesting phenomena marked in the open coding phase were grouped and their relations (causal and other) were analyzed. In the selective coding phase, a lot of data was abandoned, the core of the results - "What is going on here?" - was taken, and the research reports were written. An example of axial coding phase where i.e. associations and causal relations were searched by visualizing interesting phenomena with Atlas.ti software is presented in figure 5.

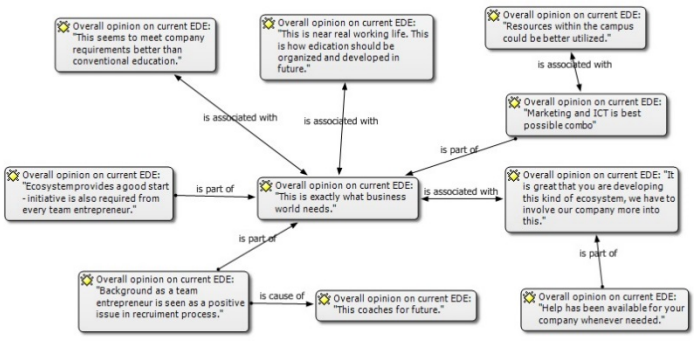


Figure 5. Axial coding with Atlas.ti code network view.

As mentioned above, the grounded theory analysis lets the data speak, and therefore no pre-existing theory is needed. In an ideal case, the grounded theory analysis is purely inductive. In practice, there always exists little or more pre-existing knowledge and bias related to the research subject. To be exact, a target to develop something is already a strong bias. Who defines development? Development for one stakeholder group may be stagnation for another group. When discussing values and biases, the best we can do as researchers is to be as open as possible about the motives we

recognize in ourselves. In this study, the researchers had a strong vision of how ICT Bachelor education should be organized in future. However, the interviewees were asked open-ended questions about the subject, and the results are presented as they are with no value-laden interpretations added.

Open coding and axial coding phases went on in parallel. During the open coding and axial coding phases, a constant comparative method and theoretical sampling [51, 52] were used. The use of the constant comparative method meant that when an interesting phenomenon was found later during data collection, all the earlier data was screened again to see if the phenomenon was found there, too. Theoretical sampling was used to collect more data on phenomena where more explanation was needed. The data was collected until a theoretical saturation was achieved. The use of the constant comparative method and theoretical sampling increased the researchers' interaction with the data. The selective coding phase also started early and parallel with the axial coding phase. The saturation of the new data [51] took place early, which meant that there was no need for extra interviews on that subject. A new tool for making word clouds was used to visualize part of the data (see Figure 4 for more details).

During the first round of interviews made with stakeholder groups, two seed categories [52] emerged during the open coding phase: 1) creating interdisciplinary interaction during ICT education, and 2) entrepreneurial mindset and its value. Based on an analysis of these seed categories and combined with the analysis of the new field notes (including memos, emails, book essays, and observation data) between January 2016 and mid November 2016, a pattern of "Where should we be heading?" was written. The results were published in a conference article in the Educon 2017 conference in Athens, Greece on 27th April 2017.

At the same time when employing designed actions to the Experimental Development Ecosystem learned from the previous study, two new research questions were chosen for this study:

1. *What are the learning experiences when dialogue is used as learning method?*
2. *How new ICT tools and applications have been utilized by team entrepreneurs during the first year?*

Between mid of November 2016 and mid November 2017, plentiful new data was gathered and analyzed. This data collection comprised two interview rounds as part of the team entrepreneurs' development discussions, over 50 training sessions (where team entrepreneurs and the team coach were present), and a survey. The survey was presented to the team entrepreneurs face to face and they were able to ask further questions about the purpose of reflecting on these issues. The team entrepreneurs were asked to evaluate their dialogue skills, make observations on dialogue in other groups they had visited, and describe how they had employed ICT tools in their project work. The expressions made by 14 interviewees were later observed in training sessions, project reflection sessions, and finally by end products made with the ICT tools that had been utilized.

4. Observations based on the data

In general, the majority of the interviewees within our stakeholder groups shared the opinion that the ecosystem model (EDE) in modern education is a response to many challenges in the employment of the newly graduated, and at the same time it is a way to pave the way to embarking the working career smoothly after studies. The students' early adoption of an open view of the real business life was generally seen as a positive and novel way of educating experts for the business world, where the described individual capabilities are appreciated by employees.

Entrepreneurial mindset and thinking as a driving force for students was seen as a positive feature to be maintained in education. The necessity of establishing a cooperative and working as team entrepreneurs was seen differently among the interviewees. Some of the interviewees from Saimaa UAS did not see that learning to run an enterprise would bring much value to ICT students. The interviewees from industry, however, saw that learning to think and operate as entrepreneurs would be one of the most important topics to learn overall.

As the most promising pathways, the interviewees saw the possibilities of interdisciplinary interaction between marketing and ICT students, who all are familiarized with the ICT business and the ICT customers' businesses. During the studies, the two student groups could make use of each other's specialty areas by participating in shared projects either in RDI or in business collaboration projects. As the study aims at understanding the future requirements for a competent ICT education, several ICT company representatives were interviewed. As a way to prepare competent future employees for the ICT field, the interviewed ICT companies mentioned a possibility to participate directly in the education of ICT students via both direct education activities and placement opportunities or shared customer projects in RDI, for example. This would create a firm an efficient recruitment process with less need for training the newly recruited personnel.

The interviewees that have graduated from the team entrepreneurship ecosystem or are currently studying marketing in that environment had positive experiences of this type of an education system. The most important and positive feature mentioned was learning through real business cases. The most valuable part in the model was the early networking with businesses. After graduation, it has been easy to find a job relating to the field of studies. At least, this different form of studies in a cooperative has proven a positively differentiating factor in job applications.

ICT business is mostly a project-based business where experience and an entrepreneurial mindset are a benefit as such. The current marketing team entrepreneurs have started to build also cooperation with students from different sectors of education within Saimaa UAS to be able to expand their domain knowledge. Making oneself familiar with different contexts where marketing activities (and ICT) are carried out is another example of a positively differentiating factor for the job applicant.

The ICT education ecosystem and ICT education in general need to respond to the changing skill requirements in the industry. In the EDE this can be promoted by involving the ICT sector in the constant development of the study programme and in the education activities in special projects. This would ideally result in long-term interaction between the ICT industry and the Saimaa University of Applied Sciences (SUAS) and the students. In some cases, even weekly co-operation between the ICT companies and the coach

and students would form a mutually beneficial cooperation model supporting a high level of motivation among both the students and the ICT companies. When mutual benefits exist and the captured value will be fairly shared between the parties, the cooperation will sustain itself.

There were some concerns towards the coached team learning and team entrepreneurship model raised in the interviews. In general, the team coaches have a key role and they are responsible for ensuring that the students are provided with the basic capabilities required in the education programme at a University of Applied Sciences. It was also stated that this type of a learning method is not ideal for students who still have to grow in order to mature for the responsibility and self-leadership skills that are required by the learning style. The entrance examination for the students was seen as one important step in this process. Also, the new ICT education should be marketed as a more interdisciplinary education program that is not too much technically oriented in order to lure the business oriented people with a high motivation for fast career building.

As the student teams operate in groups or as working teams, it would be beneficial if the different team roles could be consciously tested by each student. This would help the young students safely test their own personal strengths, which would be a supporting factor in building the students' professional self-esteem. As all companies, also the student teams have to set goals for their actions and all the steps during the studies should be taking the team towards the goals. It is not enough to act as active since that does not suffice to develop the situation. Supporting the building of students' versatile knowledge base requires a broad scale of practical learning projects, which has to be ensured during the studies. Here, the role of the team coach is emphasized again. A frequent presentation discussion of the learning goals and how they have (or have not) been achieved is crucial.

In the Business Administration degree program, the number of students who complete 55 ECTS points per year has been monitored since 2014 when criteria for funding University of Applied Sciences was changed. However, the measuring criteria were changed again in 2016, so we do not yet have enough new data to be able to make valid arguments about overall results. In the academic year 2015 – 2016, the overall result in achieving 55 ECTS points in the Business Administration and International Business degree programs (measured together) was 57.1 % (68.1 % with team entrepreneurs). In the academic year 2016-2017, the figures were 61.3 % (58.7%) respectively. The measuring system is now being updated to be able to provide more accurate data in future.

Even though we do not have enough performance data yet from team entrepreneurs' studying within the EDE, we are able to argue something. One of the biggest worries at the beginning of team learning and team entrepreneurship was how the increased freedom of students in choosing what they study and when will affect the performance measured in ECTS points. With this small amount of data we are able to argue that the pedagogics used with the EDE is working and no one has to worry about the performance metrics.

Two of the interviewees strongly emphasized that in future the EDE should emphasize the role of internationalization of students and companies cooperating in the EDE. This kind of organization of learning, development, and research activities is not familiar elsewhere and therefore the EDE ideas should be exported to other countries as well. The main strengths that this kind of ICT education would offer include the digital thinking of "Digi-native" generations together with marketing capabilities. These make a combination that a successful and competitive international business requires. The students would also, right from the beginning, start to understand that their future work will be part of someone's business – if not their own.

4.1. Towards a perfect ecosystem of learning

The key findings, as also described in pedagogic research literature that have proved to be successful in the researched EDE model are the following: authentic learning, mental model building, internal motivation, multiple intelligences as well as social learning. The curriculum applied in Saimaa UAS, as well as the piloted methodology applied on certain study courses, aims at developing skills needed when coping with the complexities of the 21st century: team working skills, communal learning skills, problem solving skills, leadership and self-leadership skills as well as innovativeness, shared expertise, and ability to reflect on one's own values and attitude.

These suggestions follow the advice from studies on how to build great teams. They suggest that energy (how team members contribute to a team as a whole), engagement (how team members communicate with one another), and exploration (how teams communicate with one another) are crucial factors for teams' success [53]. Furthermore, when quality of communication includes balance of advocacy and inquiry, a lot of positive feedback and true interest in others' opinions [54], there are many of the required elements available to develop great teams. Applications selected should support the objectives listed here, and this means they cannot be selected from a teacher-centric viewpoint.

Students' experiences from learning dialogue skills are proving that practicing dialogue in a safe environment gives them valuable communication exercise. It also supports the readiness for dialogue of the future company developers. Within one year's time every respondent said that there is development gained in one's own dialogue skills. Several respondents stated that they already are good at listening to other students. Several stated that they have learned to express their own opinions in the group. Based on external working life experiences, some students also had noticed that in many companies the leadership culture does not encourage dialogue at all. These future employees will be motivated as well as committed when they are part of a team where dialogue exists.

4.2. Where there's a will there's a way – A practical view of communication technologies and platforms.

A framework for making interpretations based on the qualitative data analysis for research question 2 had four phases:

1. Notes from interview data (What tools were said to be employed?)
2. Direct and participative observation in training sessions (What tools were actually employed in practice and by who?)

3. Reflection of what was learned (Who actually participated to implementation of new tools? How tools were selected? Why tools were abandoned?)
4. What outputs were done for customers (Videos, web-sites, etc.) with the ICT tools employed?

These phases helped to validate our findings. When a team entrepreneur expressed in an interview (1) that she was interested to learn a new tool for making videos, later discussed in training session (2) about utilizing the tool in a customer project, reflected the use of same tool when a project was completed (3), and finally the a video was publicly available (4) in social media platform – we were able to validate integrity of our chain of making interpretations. An example of how these 14 cases were analyzed is presented in figure 6.

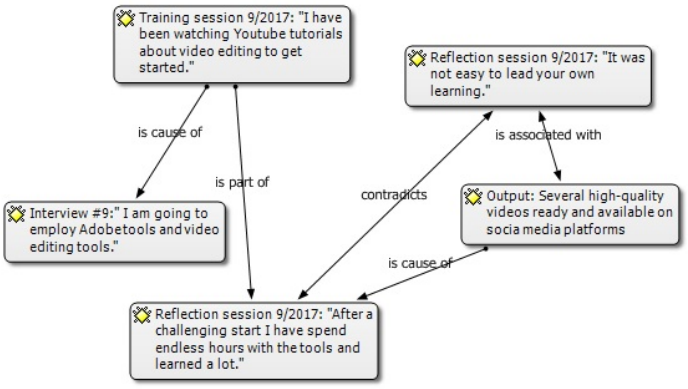


Figure 6. An example of qualitative data analysis, case #9

Based on our current experience, better utilization of tools and applications is not limited by availability of ICT tools and applications. Rather, it is limited by individual differences in the will to share information about one's working. Sharing work-related information is a much more sensitive issue than it is usually expected to be. This very same phenomenon limits the use of many other ICT systems as well. Every time when there is competition of projects within teams or departments of the same organization, openness in sharing information is limited. Only part of the employees are willing to share their business contacts via the company-wide CRM system, for example. A better understanding of the benefits of balance between individual work and team work (see [55] for more details) is needed to better utilize ICT systems. Cui bono (Who benefits?) question has to be answered clearly when new systems are employed.

Adopting the use of some practical IT tools while learning in projects is important. To facilitate communications between learners and to support project management, a broad selection of "free to use" tools have been applied within the EDE. The student respondents in the study in August 2017 stated that they had learned to use 4 - 5 new IT tools, applications or platforms on average during one year's time. Some of these tools had been presented by the team coach or a visiting company representative. Currently, also learners find and introduce new tools to the community on a regular basis. The list of the tools, applications and platforms used is long, including the following ones that were mentioned the most frequently: Trello, Slack, Skype, Doodle, WhatsApp, Moodle, Dropbox, Google Drive, Canva, Mention, Facebook, Twitter, and Instagram. Naturally, email, SMS messages and other conventional communications tools are used as well.

The most frequently used tools in project management were Google Drive, Google Sheets and Trello. In a brief unofficial discussion with the team entrepreneurs from other teams, the same tools were mentioned as the most popular ones in other teams as well.

According to our experiences, it is rather simple for the students to start using new digital tools when working on projects. This also helps students to quickly gain valuable experiences with different tools and platforms and their usability in different tasks. Sometimes students want to challenge themselves by learning to use a new tool, and harness it into the use of their cooperative. There is no need to teach the system itself, but rather act as an example and invite students to use the system. That way they will learn how the real work processes can make use of different technologies.

Though learners are using several tools and applications to support their learning in projects and to facilitate their communication, we are still in the early stages of truly benefitting from digitalization. Technology is not a limitation anymore. The real limitations are found in normal human behavior. To pick an example, a project manager and team members would benefit from knowing what tasks other team members currently have, who might need more tasks do to or who would need help. This data could easily be entered in a Trello table and shared to every team member.

4.3. Objectives first, systems to support productivity second

Making experiments and building prototypes [56, 57], associating different types of knowledge [58 - 60], and sharing knowledge of learning experiences [60 - 62] are part of the innovation management concept of the EDE. The team entrepreneurs experiment with different types of tools, applications, systems and platforms to find the ones that are appropriate and useful for their purposes.

Our target with the EDE is to help a new culture of agile methods, applications, and practices to emerge. The easiest way so far has been to start with the team entrepreneurs by introducing new applications to them and just starting using them in project management and communications, when students can directly see the benefit of the systems. So far, the results have been promising.

Experimental development is inbuilt in the EDE, in the form of innovation assignments mentioned before. Also the learning by doing approach, supported by reflection on what was learned, creates confidence in the cooperatives and students' own skills. Dedicating time for reflections is an important learning outcome on the way of becoming a professional in one's own area [63].

5. Discussion and summary

Combining and analyzing all data from the stakeholder groups of the EDE and from the team entrepreneurs studying within the ecosystem, we can clearly identify three steps we have to take next:

1) to leverage lessons learned so far in the ICT bachelor curriculum and master study curricula 2) to re-evaluate how extensively the use of new ICT tools has to be taught, and 3) to set objectives for further studies on the topic.

The current Experimental Development Ecosystem offers a pedagogically solid, tested, experimentally developed, and continuously evolving basis for supporting bachelor education in higher education. Also, the use of digital tools and applications to support learning by doing in projects and to facilitate communications within and between teams has been increasing rapidly.

5.1. Probable future for the EDE

Professional development should not focus on educating students about how current professions are executed [64]. As suggested, professional education should rather be focused on developing professions in cooperation between educational institutes and companies. Underneath these suggestions lies the activity theory [65, 66]. The ultimate goal is to help school and work life to collaborate in a better way. In times of rapid changes, educational arrangements including theory, practice and reflection repeat themselves in a continuing manner.

- International partners are needed to serve both education and business requirements. Companies who are interested in internationalizing their operations will benefit from opportunities for rapid experimentation in two or more locations.
- ICT students need a balanced amount of core ICT skills and business skills. This could be achieved by working with team entrepreneurs studying marketing with the EDE.
- ICT and marketing students should be put in mixed teams for sharing knowledge, and especially for creating new knowledge together. More cross-fertilization is needed for more learning and more versatile ideas.

There is also a lot of work to be done to better integrate the competencies of Saimaa UAS staff into the coaching process of marketing team entrepreneurs and/or ICT students, whether they study as team entrepreneurs or conventional students. This will require a more adaptive attitude towards how learning is organized from all the parties involved: teachers, administration, and team entrepreneurs.

In the beginning, this will require a lot of encouragement because even though higher education teachers may have extensive experience in teaching, they might lack competencies in learner-centric pedagogy and its methods, the mindset required for experimental development, or even both. Changing methods will also require teachers to accept that employing these methods means that they will frequently face situations that are novel to everyone. If someone has a strong routine of teaching, there might not be much interest or will to change it.

When the courage to experiment new methods to support learning is found (usually with support of colleagues or appropriate further education), at the beginning the change of habits will raise anxiety among teachers and learners. Teachers are used to acting as specialists providing answers and at the same time learners have been passive listeners. When tables are turned, both parties feel uncomfortable and the temptation to go back to old habits is strong. If this happens, it will return the potential of learners taking responsibility and teachers role of not giving easy answers to learners directly to zero. During the process of change we have gone through we have seen this phenomenon take place several times.

Very soon teachers, however, usually find out that with longer experience in life in general and operating in different types of situations helps to coach others to find means of how to solve them. Teachers are usually very good at abstracting and conceptualizing what has been learned. By reflecting on the learned skills and discussing how they can be applied in other contexts will help transfer the learning outcomes.

5.2. How teachers, lecturers and team coaches are able to benefit from the changes

At the beginning, creating a learning environment required in the 21st century will take more time than is usually used when organizing courses in higher education. After a two or three years' time, a new culture of students' responsibility for their own learning process will start to emerge and strengthen. At this tipping point, a team coach will recognize that there will be time for tactical and strategic thinking again.

5.3. Summary

By combining and analyzing the versatile data we collected, we were able to answer the research questions:

1. *What are the learning experiences when dialogue is used as learning method?*

2. *How new ICT tools and applications have been utilized by team entrepreneurs during the first year?*

This data shows that dialogue as a learning method was appropriate to support the professional development of the team entrepreneurs. Various new tools for project management, communication, and marketing and sales were employed in practice. It is noteworthy that these ICT tools, applications or systems were not taught. The dialogue and presentation of the tools was enough to inspire action.

The higher education institutes should foster creation of such ecosystems where knowledge, skills and character are combined. The learning environments should provide learner-centric methods, and acknowledge the importance of metacognition. Focusing mainly on content is not enough.

The systems (tools, applications, platforms) needed to support the ecosystem are nowadays mostly free to use and easy to use, meaning that most of them can be applied to practice at a fast pace. These are used in several visionary companies and they should be applied also in higher education institutes if they are not in use yet. The authors warmly welcome new partners to write and discuss the topics covered in this article.

Acknowledgments

This article has been written as a part of the Digikaappaus project, which is funded by Saimaa University of Applied Sciences,

Lappeenranta University of Technology, Tekes – the Finnish Funding Agency for Innovation, and by 11 Finnish organizations.

References

[1] P. Juvonen, A. Kurvinen, "Developing Experimental Development Ecosystem to serve ICT Education – A follow-up Study of Collaboration possibilites between Stakeholder Groups" in IEEE EDUCON 2017 Conference on 26th April 2017, Athens, Greece. http://ieeexplore.ieee.org/document/7943052/?reload=true

[2] C. Fadel, M. Bialik, B. Trilling, Four-dimensional-education – The competencies learners need to succeed. The Center for Curriculum Redesign, 2015.

[3] OECD, Schooling Redesigned: Towards Innovative Learning Systems, Educational Research and Innovation. OECD Publishing, 2015. http://www.keepeek.com/Digital-Asset-Management/oecd/education/schooling-redesigned_9789264245914-en#.WYvvnMuwfbg#page3

[4] S. Leppimäki, T. Meristö, H. Tuohimaa, J. Laitinen, "Future needs for competencies in ICT companies. Tulevaisuuden osaamistarpeet tietotekniikkayrityksissä", Corporate Foresight Group, CoFi / Åbo Akademi, 2007, (In Finnish).

[5] OECD, "Moving Up the Value Chain: Staying Competitive in the Global Economy", OECD, 2007.

[6] OECD (2012). Education Today 2013: The OECD Perspective, OECD Publishing.

[7] T. Meristö, S. Leppimäki, J. Laitinen, H. Tuohimaa, "The skill foresight of The Finnish Technology Industries. Tulevaisuuden osaamistarpeet teknologiateollisuudessa Yhteenvetoraportti toimialakohtaisista yrityskyselyistä.", Teknologiateollisuus ry, 2008. (In Finnish).

[8] B. Trilling, C Fadel, 21st Century Skills – Learning for Life in our Times, Jossey-Bass, 2009.

[9] P. Ylä-Anttila, "Dispersion of value chains, future skills requirements change. Sähkö-, elektroniikka- ja tietotekniikka-ala. Tuotantoketjut hajautuvat, osaamistarpeet muuttuvat." ETLA, The Research Institute of the Finnish Economy. 2012, (In Finnish).

[10] R. Rajander-Juusti, "Competency needs in Business and Administration. Liiketalouden osaamistarpeet. Ennakkotietoa koulutuksen suunnittelun tueksi." Raportit ja selvitykset 2013:1. Opetushallitus, 2013, (In Finnish).

[11] D. Bohm, On dialogue, Routledge Classics, 1996.

[12] W. Isaacs, Dialogue: The art of thinking together. Doubleday, Randomhouse Inc, 1999.

[13] Tiimiakatemia. Visited on 11/2017, http://tiimiakatemia.fi/en

[14] Proakatemia. Visited on 11/2017, http://proakatemia.fi/en/

[15] P. Juvonen, "LEARNING INFORMATION TECHNOLOGY BUSINESS IN A CHANGING INDUSTRY LANDSCAPE. The Case of Introducing Team Entrepreneurship in Renewing Bachelor Education in Information Technology in a University of Applied Sciences." Doctoral Thesis, Acta Universitatis Lappeenrantaensis 606, 2014.

[16] V. A. Alexandria, P. Parton, A. Robb, "Entrepreneurship Education and Training Programs around the World. Dimensions for Success.", International Bank for Reconstruction and Development / The World Bank, 2014.

[17] R.A. Baron, "Behavioural and Cognitive Factors in Entrepreneurship: Entrepreneurs as the Active Element in New Venture Creation", Strategic Entrepreneurship Journal **1**, 167–82, 2007.

[18] C. K. Volkmann, D. B. Audretsch (Eds.) "Entrepreneurship Education at Universities. Learning from Twenty European Cases", Springer International Publishing 2017.

[19] D. La Guardia, M. Gentile, V. Dal Grande, S. Ottaviano, M. Allegra, "A Game based Learning Model for Entrepreneurship Education", Procedia - Social and Behavioral Sciences, **141**, 195-199, 2014.

[20] A. Daniel, R. Costa, M. Pita, C. Costa, "Tourism Education: What about entrepreneurial skills?", Journal of Hospitality and Tourism Management, **30**, 65-72, 2017.

[21] R.G. Klapper, V.A. Farber, "In Alain Gibb's footsteps: Evaluating alternative approaches to sustainable enterprise education (SEE)", The International Journal of Management Education, **14**(3), 422-439, 2016.

[22] R. Bell, "Developing the next generation of entrepreneurs: Giving students the opportunity to gain experience and thrive", The International Journal of Management Education, **13**(1), 37-47, 2015

[23] S. Mat, S. Maat, N. Mohd, "Identifying Factors that Affecting the Entrepreneurial Intention among Engineering Technology Students", Procedia - Social and Behavioral Sciences, **211**, 1016-1022, 2015.

[24] C. Shlaegel, M Koenig, "Determinants of Entrepreneurial Intent: A Meta-Analytic Test and Integration of Competing Models", Entrepreneurship Theory and Practice, **38**(2), 291-332, 2013.

[25] C. Luthje, N. Franke, "The "making" of an entrepreneur: Testing a model of entrepreneurial intent among engineering students at MIT", R&D Management, **33**(2), 135 – 147, 2003.

[26] B. Hj Din, A. Rahim, R. Anuar, M. Usman, "The Effectiveness of the Entrepreneurship Education Program in Upgrading Entrepreneurial Skills among Public University Students", Procedia - Social and Behavioral Sciences, **224**, 117-123, 2016.

[27] B, Askun, N. Yıldırım, "Insights on Entrepreneurship Education In Public Universities In Turkey: Creating Entrepreneurs Or Not?", Procedia - Social and Behavioral Sciences, **24**, 663-676, 2011.

[28] J. Kirkwood, K. Dwyer, B. Gray, "Students' reflections on the value of an entrepreneurship education", The International Journal of Management Education, **12**(3), 307-316, 2014.

[29] A.S. Zamberi, A. Bakar, N. Ahmad, "An evaluation of teaching methods of entrepreneurship in hospitality and tourism programs", The International Journal of Management Education, **16**(1), 14-25. Forthcoming, 2018.

[30] M. Küttim, M. Kallaste, U. Venesaar, A. Kiis, "Entrepreneurship Education at University Level and Students' ", Entrepreneurial Intentions, Procedia - Social and Behavioral Sciences, **110**, 658-668, 2014.

[31] T. Shih, Y-Y. Huang, "A case study on technology entrepreneurship education at a Taiwanese research university." Asia Pacific Management Review, **22**(4), 202-211, 2017. https://doi.org/10.1016/j.apmrv.2017.07.009

[32] I, S, Ahmad, M, F. R. Buchanan, "Examining the entrepreneurship curriculum in Malaysian polytechnics." The International Journal of Management Education, **12**(3), 397-406, 2014.

[33] S , Jansen, T. De Zande, S. Brinkkemper, E. Stam, V. Varma, "How education, stimulation, and incubation encourage student entrepreneurship: Observations from MIT, IIIT, and Utrecht University", The International Journal of Management Education, **13**(2) 170-181, 2015.

[34] J. Mezirow, Transformative dimensions of adult learning, Jossey-Bass, 1991.

[35] R. Kiely, "A Transformative Learning Model for Service-Learning: A Longitudinal Case Study", Michigan Journal of Community Service Learning. **12**(1), 5-22, 2005.

[36] P. Juvonen, "Comparison of two learning and team entrepreneurship models at a Finnish University of Applied Sciences. Setting the scene for future development", **7**(1). International Journal of Engineering Pedagogy, 2017. http://online-journals.org/index.php/i-jep/article/view/6517

[37] D. Kolb, D, Experiental Learning, Prentice-Hall, Englewood Cliffs, NJ, 1984

[38] L. B. Resnick, Introduction. In L. B. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser, 1-24, Hillsdale, NJ: Erlbaum, 1989.

[39] V. Richardson, "Constructivist Pedagogy." University of Michigan. Teachers College Record, 105(9) 1623-1640 , 2003. Teachers College, Columbia University, 0161-4681.

[40] R. Sarder, Building an Innovative Learning Organization. A framework to build a smarter workforce, adapt to change and drive growth, Wiley, 2006.

[41] D. Levi, Group Dynamics for teams, 2nd edition, Sage Publications, 2007.

[42] M. Beer, High Commitment High Performance. How to Build a Resilient Organization for Sustained Advantage, Jossey-Bass, 2009.

[43] P. Cockman, B. Evans, B. P. Reynolds, Consulting for Real People. Second Edition. McGraw-Hill Publishing Company, 1999.

[43] K. Herr, G.L.Anderson, The Action Research Dissertation. A guide for Students and Faculty, Sage Publications, 2005.

[45] E.T. Stringer, Action Research 3rd ed, Sage Publications, 2007.

[46] R.K. Yin, Case Study Research: Design and Methods, Sage, 2003.

[47] K.M. Eisenhardt, M.E. Graebner, Theory building from cases: Opportunities and challenges, Academy of Management Journal, **50**(1) 25 – 32, 2007.

[48] N.K. Denzin, Y.S. Lincoln, (eds.), Collecting and Interpreting Qualitative Materials, Sage Publications 2003.

[49] A. Strauss, J. Corbin, Grounded Theory in Practice, Sage Publications, 1997.

[50] U. Flick, An Introduction to Qualitative Research. 4th ed. Sage Publications, 2009.

[51] B.G. Glaser, Theoretical sensitivity. Advances in the Methodology of Grounded Theory. Sociology Press, 1978.

[52] B.G. Glaser, Emergence vs. Forcing: basics of grounded theory analysis. Sociology Press, 1992.

[53] A. Pentland, The New Science of Building Great Teams. Harvard Business Review, April 2012.

[54] M. Losada, E. Heaphy, The Role of Positivity and Connectivity in the Performance of Business Teams. American Behavioral Scientist, **47**(6), 740 – 765, 2004.

[55] P. Lencioni, The five dysfunctions of a Team. A workshop for teams, Jossey-Bass, 2012.

[56] M. Schrage, Serious Play. How the World's Best Companies Simulate to Innovate, Harvard Business School Press, 2000.

[57] N. Furr, Dyer, J, The Innovator's Method. Bringing the Lean Startup into your Organization, Harvard Business Review Press, 2014.

[58] J. Dyer, H. Gregersen, C.L. Christensen, The Innovator's DNA. Mastering the Five Skills of Disruptive Innovators. Harvard Business Review Press, 2011.

[59] E. Wenger, R. McDermott, W.M. Snyder, Cultivating Communities of Practice. A guide to managing knowledge, Harvard Business School Press 2002.

[60] J.R. Katzenbach, D.K. Smith, The Discipline of Teams. A Mindbook-Workbook for Delivering Small Group Performance, John Wiley & Sons, 2001.

[61] G. Von Krogh, K. Ichijo, I. Nonaka, Enabling Knowledge Creation. How to Unlock the Mystery of Tacit Knowledge and Release the Power of Innovation, Oxford University Press, 2000.

[62] P.J. Denning, R. Dunham, The Innovators Way. Essential practices for successful Innovation, The MIT Press, 2010.

[63] D. Boyd, R. Keogh, D. Walker, Reflection: Turning Experience into Learning, Kogan Page, 1985.

[64] J. Virkkunen, H. Ahonen, M. Schaupp, L. Lintula, Toimintakonseptin yhteisen kehittämisen mahdollisuus. TYKES, raportteja 70, 2010, (In Finnish).

[65] L.S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, 1978.

[66] Y. Engeström, Developmental work research: Expanding activity theory in practice, Lehmanns Media, 2005.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 94-102 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Auto-Encoder based Deep Learning for Surface Electromyography Signal Processing

Marwa Farouk Ibrahim Ibrahim*, Adel Ali Al-Jumaily

School of Mechanical and Electrical Engineering, Faculty of Engineering and IT, University of Technology Sydney, NSW 2007, Australia

ARTICLE INFO

Article history:
Received: 01 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords:
Feature learning
Deep learning
Bio-signal processing
Sparse autoencoder
Analysis of variance
Confidence interval
Support vector machine
Extreme learning machine
Softmax layer
Classifier fusion

ABSTRACT

Feature extraction is taking a very vital and essential part of bio-signal processing. We need to choose one of two paths to identify and select features in any system. The most popular track is engineering handcrafted, which mainly depends on the user experience and the field of application. While the other path is feature learning, which depends on training the system on recognising and picking the best features that match the application. The main concept of feature learning is to create a model that is expected to be able to learn the best features without any human intervention instead of recourse the traditional methods for feature extraction or reduction and avoid dealing with feature extraction that depends on researcher experience. In this paper, Auto-Encoder will be utilised as a feature learning algorithm to practice the recommended model to excerpt the useful features from the surface electromyography signal. Deep learning method will be suggested by using Auto-Encoder to learn features. Wavelet Packet, Spectrogram, and Wavelet will be employed to represent the surface electromyography signal in our recommended model. Then, the newly represented bio-signal will be fed to stacked autoencoder (2 stages) to learn features and finally, the behaviour of the proposed algorithm will be estimated by hiring different classifiers such as Extreme Learning Machine, Support Vector Machine, and SoftMax Layer. The Rectified Linear Unit (ReLU) will be created as an activation function for extreme learning machine classifier besides existing functions such as sigmoid and radial basis function. ReLU will show a better classification ability than sigmoid and Radial basis function (RBF) for wavelet, Wavelet scale 5 and wavelet packet signal representations implemented techniques. ReLU will illustrate better classification ability, as an activation function, than sigmoid and poorer than RBF for spectrogram signal representation. Both confidence interval and Analysis of Variance will be estimated for different classifiers. Classifier fusion layer will be implemented to glean the classifier which will progress the best accuracies' values for both testing and training to develop the results. Classifier fusion layer brought an encouraging value for both accuracies either training or testing ones.

1. Introduction

Supervised learning is widely utilised in various applications. However, it is still quite limited method. The majority of applications need handcrafted engineering extraction of features by implementing different techniques. This means that the principal purpose is to represent the bio-signal by applying proper feature representation methods. Whenever significant features represent bio-signal, classification error should be anticipated to be lower than extracting features, which are not genuinely representing data. However, the general engineering handcrafted representation is still effortful and consumes a long time. Moreover, the standard feature extraction algorithm relies on researcher's experience. Many proposed feature learning methods may be implemented to improve feature representation automatically and save both effort and time. The primary evaluation of the behaviour of implemented feature learning method is the classification error. Deep learning is considered the

*Marwa Farouk Ibrahim Ibrahim, 15 Broadway, Ultimo, NSW 2007, Australia, 0061405565739 & marwafaroukibrahim.ibrahim@student.uts.edu.au

most common technique to implement feature learning. Rina Detcher was the first to introduce the fundamentals for both first and second order deep learning [1]. Deep learning is an essential division of machine learning that consists of a multilayer. The output of each layer is considered as features that will be introduced to the following cascaded layer [2]. Artificial neural networks use a hidden layer to implement each layer of multilayers that construct deep learning [3]. Fig.1 shows a simple architecture of deep learning steps. The learning technique is done in a hierarchal method starting from the lower layers to the upper ones [4]. Deep learning can be used for both supervised and unsupervised learning where it learns features from data and eliminates any redundancy that might be existing in the representation. Unsupervised learning recruitment brings more defy than supervised one. Unsupervised learning for deep learning was implemented by Neural history compressors [5]and deep belief networks[6].

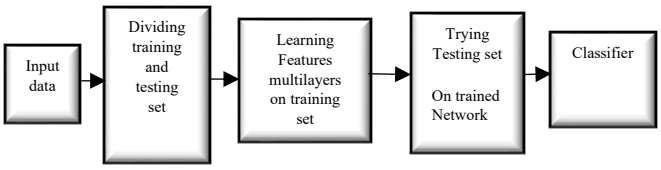


Fig.1. Traditional Deep Learning Steps

This paper is organised by presenting a brief study on previous work that has been done on classification finger movement and deep learning in different fields, then a review study on autoencoder including the main equation for Auto-Encoder will be introduced. The surface electromyography will be assimilated by Wavelet Packet, Spectrogram and Wavelet. We will compare our results by implementing three different classifiers, which will be Support vector machine, Extreme learning machine with three activation functions and Softmax layer.

The Analysis of Variance (ANOVA) will be calculated for different classifiers in Auto-Encoder deep learning method. Also, the confidence interval for Auto-Encoder will be implemented as well. At last, each of training and testing accuracy will be promoted by concatenating classifier fusion layer.

2. Previous Work

In this research, we will suggest a deep learning system that will be capable of providing essential features from the input signal without recourse to traditional feature extraction and reduction algorithms. The suggested system will be talented in assert the ten hand finger motions. The classification of different Finger motions was discussed earlier in many published scientific types of research. The early pattern recognition for finger movements was proposed in [7] where the researchers suggested using neural networks in analysing and classifying the introduced EMG pattern. They classified both finger movement and joint angle associated with moving finger. Later, in [8] the authors investigated and optimised configuration between electrode size and its arrangement to achieve high classification accuracy. Then, in [9] the researchers gave more attention to selecting the extremely discriminative features by employing Fuzzy Neighbourhood Preserving Analysis (FNPA) where the main purpose of this technique is to reduce the distance between the samples that belong to the same class and maximise it between samples of different classes. In the same year, other researchers explored the traditional machine learning well-known algorithm. Where, they used time domain features and implemented support vector machine, linear discriminate analysis and k-nearest neighbours as different classifiers then, they took advantage of Genetic Algorithm to search for redundancy in the used dataset and selected features as well [10]. In the same context, authors proposed an accurate finger movement classification system by extracting time domain-auto regression features, reducing features by using orthogonal fuzzy neighbourhood discriminant analysis technique and implementing linear discriminant analysis as classifier [11]. After that, other researchers suggested an accurate pattern recognition system for finger movement by extracting 16-time domain features to process the Electromyography signal and implementing two layers feed forward neural networks as classifiers [12]. In contrast, effort and time that are being wasted, as mentioned before, in feature extraction and reduction were the motivation behind introducing the concept of deep learning. Therefore, many researchers published valuable achievements in deep learning for the biomedical signal. An extensive review study was presented on different types of research that recalled deep learning in health field [13]. The common factor in each study was the recruitment of neural network to learn features from input bio-signal. In the same context, researchers proposed a model by using convolutional neural networks to convert the information which was given by wearable sensor into highly related discriminative features [14]. Another research presented a deep learning record system that predicted the future medical risk automatically after extracting essential features by implementing convolutional neural networks [15]. Also, researchers implemented a system that used to extract shallow features from wearable sensor devices then the features were introduced to convolutional neural networks and finally to the classifier layer [16]. Based on the above, we can conclude that deep learning is an initial step towards implementing self-learning system by using neural networks. In our proposed system, we will implement neural networks in the form of two stages autoencoder, which read represented bio-signal by either spectrogram, wavelet or wavelet packet. We will use different classifiers to evaluate our system behaviour. Finally, we will add classifier fusion layer, which will follow best local classifier methodology. Adding classifier fusion was a promising contribution to the accuracies. Moreover, both confidence interval and Analysis of Variance will be estimated for different classifiers.

3. Sparse Auto-Encoder

An Autoencoder is an extensively used technique to reduce dimensions [17]. Sparse autoencoder idea first started in [18]. Where it started to reduce the redundancy that may result from complex statistical dependencies. Building a neural network and train it by using sparse method penalty as mentioned in [19] and taking into account the number of hidden nodes in the developed neural network, is considered a straightforward factor but as crucial as choosing the learning algorithm [20].

Auto-Encoder is a feed-forward neural network that is used in unsupervised learning [21]. The implemented neural network is being trained to learn features and produce it as its output rather than generating classes in case of recalling the classification ability of the hired neural network [22]. The encoder input is the represented data while its output is the features learnt by autoencoder. The learnt features learnt from the autoencoder will be introduced to classifier to be used in the assort of the data into

predefined classes[23]. Lately, autoencoder is commonly employed to extract highly expressing features from data.

Unlabelled data can be used to train an autoencoder where training is mainly interested in optimising the cost function. The cost function is mainly responsible for estimating the miscalculation that may occur in calculating the reconstructed copy of input at the output and the input data.

Assume that we have an input vector $x \in R^{D_x}$. The autoencoder maps this input to a new vector $z \in R^{D^{(1)}}$.

$$z^{(1)} = h^{(1)}(W^{(1)}x + b^{(1)}) \quad (1)$$

Where the superscript (1) represents the first layer of the autoencoder. $h^{(1)}: R^{D^{(1)}} \rightarrow R^{D^{(1)}}$ represents the transfer function, $W^{(1)} \in R^{D^{(1)}}$ represents the weight matrix, and $b^{(1)} \in R^{D^{(1)}}$ represents the bias vector. Then the decoder transfers the encoded representation z as a reconstruction of the input x following the next equation

$$\hat{x} = h^{(2)}(W^{(2)}x + b^{(2)}) \quad (2)$$

Where the upper character (2) signifies the second layer of the autoencoder. $h^{(2)}: R^{D^{(2)}} \rightarrow R^{D^{(2)}}$ accounts for the transfer function, $W^{(2)} \in R^{D^{(2)}}$ represents the weight matrix, and $b^{(2)} \in R^{D^{(2)}}$ represents the bias vector.

The sparsity term can be introduced to autoencoder by adding an adapted cost function in the form of regularisation term. The regularisation function is estimated for each neuron i by averaging its activation function, which can be expressed as follows

$$\hat{\rho}_i = \frac{1}{n}\sum_{j=1}^{n} z_i^{(1)}(x_j) = \frac{1}{n}\sum_{j=1}^{n} h(W_i^{(1)T}x_j + b_i^{(1)}) \quad (3)$$

Where n is the number of training samples, x_j is the j^{th} training sample of input, $W_i^{(1)T}$ is the i^{th} row of the weight matrix transpose of the first layer, and $b_i^{(1)}$ is the i^{th} term of the bias vector for the neural network. The neurone is considered to be firing if its output activation function is high and in the case of having a low activation value, this means that the neurone is only responding to a small number of input samples, which in turn encourages the autoencoder to learn. Accordingly, adding a limitation term to activation function output $\hat{\rho}_i$ limits every neurone to learn from limited features. This motivates the other neurones to respond to only another small number of features, which initiates every neurone to be responsible for responding to individual features for each input.

Introducing a sparsity regularise value is considered as a measure of how far or close is the targeted activation value ρ from the actual activation output function $\hat{\rho}$. Kullback-Leibler divergence is a very well know the equation that describes the difference between two different distributions. This equation is shown as follows:

$$\Omega_{Sparsity} = \sum_{i=1}^{D^{(1)}} KL(\rho||\hat{\rho}_i) = \sum_{i=1}^{D^{(1)}} \rho \log\frac{\rho}{\hat{\rho}_i} + (1-\rho)\log(\frac{1-\rho}{1-\hat{\rho}_i}) \quad (4)$$

The cost function is decreased to initiate the two distributions $\hat{\rho}_i$ and ρ to be as close as possible. The cost function can be represented by a mean square error equation as follows:

$$E = \underbrace{\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}(X_{kn} + \hat{X}_{kn})^2}_{\text{Mean Square Error}} + \underbrace{\lambda * \Omega_{weights}}_{\text{L2 Regularisation}} + \underbrace{\beta * \Omega_{Sparsity}}_{\text{Sparsity Regularisation}} \quad (5)$$

Where L2 regularisation is a term to be added to the cost function to regulate and prevent the value of Sparsity Regularisation value of being small during the training due to the increase that may happen to the values of weights and decrease to the value of the mapped vector z

$$\Omega_{wights} = \frac{1}{2}\sum_{l}^{L}\sum_{j}^{n}\sum_{i}^{K}(w_{ji}^{(1)})^2 \quad (6)$$

L is the number of hidden layers, N is the number of input data samples, and K is the number of classes.

Autoencoder was hired in many research areas as a feature learning layer. Where its primary task, was to learn features from input data. A robust study was published to compare between many applications for autoencoder in deep learning field [24]. Autoencoder was implemented in [25] to learn incremental feature learning by introducing an extensive data set to denoising autoencoder. Denoising autoencoder provides an extremely robust performance against noisy data with a high classification accuracy [26, 27]. Another suggested autoencoder was a marginalised stacked one which showed a better performance, with high dimensional data, than the traditionally stacked autoencoder regarding accuracy and simulation time [28]. Denoising stacked autoencoder was hired to learn features from unlabeled data in a hierarchical behaviour [29] and was applied to filter spam by following greedy layer-wise to the implemented denoising stacked autoencoder [30].

In our proposed model, Auto-Encoder is a feed-forward neural network that is used in feature learning. The implemented neural network is being trained to learn features and produce it as its output rather than generating classes in case of recalling the classification ability of the hired neural network. Where, we implemented a stack autoencoder, which consists of two successive encoder stages. The input to the encoder is the data while the output is the features or representations. The classifier uses features produced from encoder as an input while; its output is the classes equivalent to input data [23]. Fig.2 demonstrates the steps that the surface electromyography signal moves through by using a sparse autoencoder.

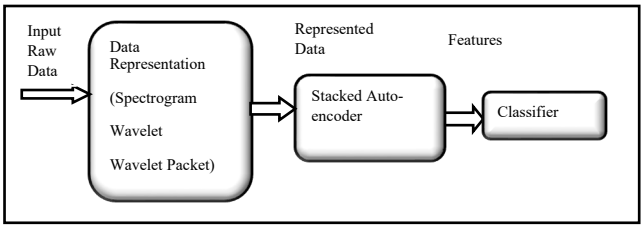


Fig.2. Procedures of Sparse Auto-Encoder signal representation

In the same context of feature learning, autoencoder will generate useful features at the output, rather than producing classes, by decreasing the dimension of the input data into a lower dimension. However, the new lower dimension data will be dealt with our features that contain essential and discriminative information on the data, which will help in better classification results. Sparse autoencoder enhances us to leverage the availability of data.

4. Bio Signal Representation

We suggested three signal representations be applied on raw biological data to ensure fidelity and precision of our bio signal. Moreover, introducing raw data directly to first auto-encoder stage resulted in accuracy less than 50%. The first data representation was the spectrogram for bio raw signal. The spectrogram is interpreted to be the illustration of the spectrum of frequencies of our surface electromyography signal in a visible method. Numerically, Spectrogram can be estimated by calculating the square of the magnitude of Short-Time Fourier Transform (STFT). It can be called short-term Fourier transforms rather than spectrogram. In short time Fourier transforms, the long-time signal is divided into equal length segments and shorter in time. Short time Fourier transforms is relevant to Fourier transform. Then, the frequency and phase for each segment to be estimated separately. Based on the above, we can deduce that spectrogram can be treated as Fourier transform but for shorter segments rather than estimating it from the full long signal at once[31].

Assume that we have a discrete time signal x with a finite duration (limited signal) and a number of samples N. The Discrete Fourier Transform (DFT) can be expressed as follows:

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n) e^{-i\frac{2\pi k}{N}n} \quad , k = 0, \dots\dots, N-1 \qquad (7)$$

Knowing that the Fourier transform is estimated at frequency $f = \frac{k}{N}$

The original signal x can be restored back from $\hat{x}$ by applying the inverse Discrete Fourier Transform as follows:

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} \hat{x}(k) e^{-i\frac{2\pi k}{N}n}, \qquad n = 0, \dots\dots, N-1 \qquad (8)$$

The above-mentioned two equations can be rephrased as follows:

$$x = \frac{1}{N} F\hat{x} \qquad (9)$$

$$\hat{x} = \bar{F}x \qquad (10)$$

Where F is Fourier matrix of $n * n$ dimensions and $\bar{F}$ is its complex conjugate

$$F = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{i\frac{2\pi}{N}} & e^{i\frac{4\pi}{N}} & \cdots & e^{i\frac{2\pi(N-1)}{N}} \\ 1 & e^{i\frac{4\pi}{N}} & e^{i\frac{8\pi}{N}} & \cdots & e^{i\frac{2\pi 2(N-1)}{N}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{i\frac{2\pi(N-1)}{N}} & e^{i\frac{2\pi 2(N-1)}{N}} & \cdots & e^{i\frac{2\pi(N-1)^2}{N}} \end{bmatrix} \qquad (11)$$

Where the entries of $\hat{x}$ is expressed in terms frequencies coefficients

$$f = 0, {}^{1}/_{N}, {}^{2}/_{N}, \dots\dots {}^{(N-1)}/_{N} \qquad (12)$$

We need to calculate the spectrogram of the signal. Assume that we have a signal x of length N, which is divided into successive equal segments m. Where $m < N$. The matrix of successive equal segments can be expressed as X where $X \in R^{mx(N-m+1)}$. The first column of X matrix equals $[x[0], x[1], \dots\dots x[m-1]]^T$ and its second column equals $[x[1], x[2], \dots\dots x[m]]^T$. The spectrogram for a signal x with window size m can be annotated $\widehat{X}$. The columns which are composing matrix $\hat{X}$ is the discrete Fourier transform

$$\hat{X} = \bar{F}X \qquad (13)$$

$$X = \frac{1}{m} F\hat{X} \qquad (14)$$

The rows of the matrix $\hat{X}$ are representing the signal x in the time domain while its columns are representing the signal x in the frequency domain. So simply spectrogram is a time-frequency representation of the signal x.

The spectrogram was used in many applications especially for speech signal analysis wherein [32] the authors represented the speech signal by different representations like Fourier and spectrogram to conclude that the resolution is mainly dependent on used representations. In [33] the researchers estimated the time corrected version of rapid frequency spectrogram of the speech signal which showed a better ability to follow the alterations in the bio-signal than other published techniques.

The second signal representation used was wavelet of the signal. Wavelet is estimated by shifting and scaling small segmentations of the bio-signal. Fourier transform is an illustration of the signal in a sinusoidal wave by using various frequencies while wavelet is the illustration of the abrupt changes that happen to the signal. Fourier transform is considered a good representation of the signal in case of having a smooth signal. While wavelet is believed to be a better representation, than Fourier, for the sudden changing signal. Wavelet gives the opportunity to represent rapid variations of the signal and help the system extract more discriminative features. We implemented Haar wavelet for our proposed model.

So in brief, a wavelet is an analysis for time series signal that has non-stationary power at many frequencies [34]. Assume that we have a time series signal x_n with equal time spacing δt and $n = 0, \dots\dots N-1$ where the wavelet function is $\Psi_o(\eta)$ that depends on time η. The wavelet signal has zero mean and is represented in both time and frequency domain [35]. Morlet wavelet can be estimated by modulating our time domain signal by Gaussian as follows:

$$\Psi_o(\eta) = \pi^{\frac{-1}{4}} e^{i\eta\omega_o} e^{\frac{-\eta^2}{2}} \qquad (15)$$

Where ω_o is the frequency of the unmodulated signal. The continuous wavelet of a discrete signal x_n is the convolutional of x_n with scaled and shifted version of $\Psi_o(\eta)$

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'}\ \Psi^*\left[\frac{(n'-n)\delta t}{s}\right] \qquad (16)$$

Where * is the complex conjugate and s is the scale.

The wavelet transform was applied in several studies and different fields as in [36] where wavelet implemented in Geophysics field, in [37, 38] for climate, in [39] for weather, in [40] and many other applications. The above equation can be simplified by reducing the number of N. The convolutional theorem permits to estimate N convolutional in Fourier domain by implementing Discrete Fourier Transform (DFT). The Discrete Fourier Transform for x_n.

$$\hat{x}_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{-2\pi i k n/N} \quad (17)$$

Where $k = 0, \ldots, N-1$ which is representing the frequencies. For a continuous signal $\Psi(t/s)$ is defined as $\hat{\psi}(s(\omega)$. Based on the convolutional theorem, the inverse Fourier transform is equal to wavelet transform as follows:

$$\psi_n(s) = \sum_{k=0}^{N-1} \hat{x}_k \hat{\psi} * (s\omega_k) e^{i\omega_k n \delta t} \quad (18)$$

Where the angular frequency ω_k can be expressed as follows:

$$\omega_k = \begin{cases} \frac{2\pi k}{N\delta t} & : k \le \frac{N}{2} \text{ or} \\ \frac{-2\pi k}{N\delta t} & : k > \frac{N}{2} \end{cases} \quad (19)$$

An improved copy of wavelet algorithm was recalled in [41] where the authors presented two techniques. The first one used expansion factors for filtering while the other one is factoring wavelet transform. The researchers in [42] introduced the Morlet wavelet to vibration signal of a machine. The vibration signal of the low signal to noise ratio was represented by wavelet to grant fidelity to the signal and allow extraction better powerful features. This model was implemented in [43] where researchers used wavelet transform to predict early malfunction symptoms that may happen in the gearbox.

As a refinement act, we scaled the wavelet signal by five in wavelet, which in turn promoted the results as shown in Table I. As a comparative study, we utilised wavelet packet for the signal representation. The signal can be represented in both time and frequency domain simultaneously. This representation gains a fidelity to the signal due to its robust representation. Wavelet packet is one of the very widely used signal representation that produces the signal in both time and frequency domain [44]. The wavelet packet shows a very well acted for both nonstationary and transient signals [45-47]. Wavelet packet is estimated by a linear combination of wavelets. The coefficients of linear combination are calculated by recursive algorithm [48]. The wavelet packet estimation can be done as follows:

Assume that we have two wavelets type signal $h(n)$, $g(n)$ and two filters of length $2N$. Let us assume that the following sequence of functions is representing wavelet functions.

$$(W_n(x), n = 0,1,2, \ldots\ldots)$$

$$W_{2n}(x) = \sqrt{2}\sum_{k=0}^{2N-1} h(k) W_n(2x-k)$$

$$W_{2n+1}(x) = \sqrt{2}\sum_{k=0}^{2N-1} g(k) W_n(2x-k) \quad (20)$$

Where $W_0(x) = \varphi(x)$ is the scaling function, $W_1(x) = \omega(x)$ is the wavelet function.

$N = 1, h(0) = h(1) = \frac{1}{\sqrt{2}}$ and $g(0) = -g(1) = \frac{1}{\sqrt{2}}$

The above equations became

$$W_{2n}(x) = W_n(2x) + W_n(2x-1)$$

And

$$W_{2n+1}(x) = W_n(2x) + W_n(2x-1) \quad (21)$$

Where $W_0(x) = \varphi(x)$ is the scaling function, $W_1(x) = \omega(x)$ is the Haar wavelet function.

Many researchers implemented wavelet packet as in [49]. The authors used wavelet packet to create an index called rate index to detect the damage that may happen to the structure of any beam. In the same context authors of [50] employed wavelet packet and neural networks to detect a fault in a combustion engine. The implemented wavelet packet was six levels for sym10 at sampling frequency 2 kHz.

5. Classifiers

In the implementation of Auto-Encoder as feature learning algorithm, we applied three different classifiers, where the first was Softmax layer, the second was Extreme learning machine, and the third was Support Vector Machine (SVM). We measured the accuracy of Linear support vector machine, Quad support vector machine, Cubic support vector machine, Fine Gauss support vector machine, Medium Gauss support vector machine and Coarse Gauss support vector machine and elected the support vector machine classifier that resulted in the highest accurate result. Furthermore, the appending of classifier fusion layer to nominate best local classifier which in return endorsed the accuracy values. Fig.3 shows the block diagram for our implemented autoencoder feature learning proposed model

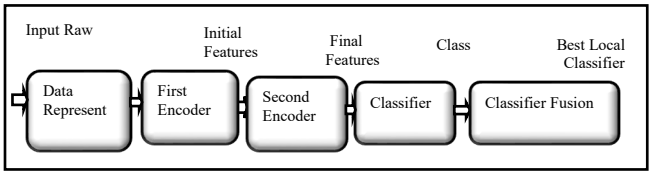


Fig.3. Scheme of proposed Model

Moreover, ANOVA for autoencoder different classifiers was implemented. Where, we assembled average testing accuracies for four signal representation techniques (Wavelet, Wavelet Scale5, Wavelet Packet and Spectrogram) that resulted in P value 0.7487. So as wavelet results should not be counted, due to its low accuracy values, so, we suggested a second trial which was to group average testing accuracies for three signal representing techniques (Wavelet Scale5, Wavelet Packet and Spectrogram) that resulted in P value 0.3405. Both P values showed that there was no sensible variation between any of the implemented three

classifiers as P value was higher than 0.05 in both cases. Fig.7 shows different P values for different classifiers.

In addition to the above, we estimated the confidence interval for each classifier. Our confidence interval was designed for confidence score 60%. Our assessed interval was bounded by higher and lower limit. In other words, we were confident or assured of any new accuracy by percentage 60% as long as it is located in the previously estimated interval.

6. Implementation

In this part, the data acquisition methods we followed will be expressed more extensively, and simulation outcomes will be exhibited and discussed.

6.1. Data Acquisition

The surface Electromyography signal was read by using FlexComp Infiniti™ device. Two sensors were placed on the forearm of the participant of type T9503M. The placement of two electrodes on participant's forearm is as shown in fig.4

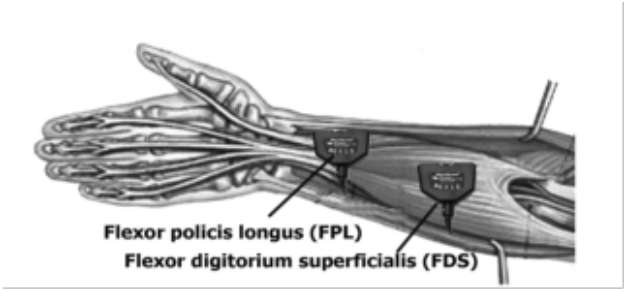


Fig.4. Placement of the electrodes

The Electromyography signal was collected from nine participants. Each participant performed one finger movement for five seconds then had a rest for another five seconds. Every finger motion was reiterated six times. The same sequence was repeated for the second finger activity. Amplification of the signal by 1000 was applied and a sampling rate of 2000 samples for each second was implemented.

The collected Electromyography signal was used to categorise between predefined ten finger motions, as shown in Fig.5 via using our suggested model. Three folded cross validation was applied on our collected Electromyography signal. Accordingly, 2/3 of the collected data was assigned to training set while remaining 1/3 to be used by testing set.

Our surface electromyography signal was filtered to ensure fidelity and removal of any noise that may be inserted into the collected bio-signal. The average training or testing accuracy was estimated by simulating our proposed model for every subject apart then summed the accuracies for all subjects and divided the result by the total number of subjects.

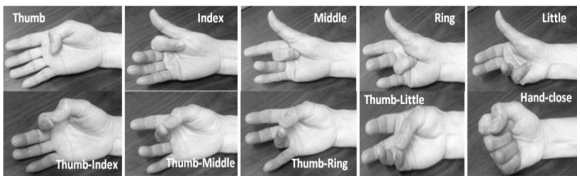


Fig.5. Targeted Ten different finger motions

6.2. Results

We implemented 400 nodes for the first layer of autoencoder and 300 nodes for the second one. As for the transfer function of the encoder, it was the pure linear type.

Table I shows autoencoder feature learning testing and training accuracy where the bio-signal was presented by spectrogram, wavelet, Wavelet scale 5 and wavelet packet. Three different classifiers were executed. The first was a SoftMax layer. While the second classifier was an extreme learning machine, we examined the performance of various activation functions for extreme learning machine classifier like sigmoid, the rectified linear unit and radial basis function. As that, the third classifier was support vector machine. We examined the performance of linear support vector machine, quadratic support vector machine, cubic support vector machine, fine gauss support vector machine, medium gauss support vector machine and coarse gauss support vector machine. Then, we selected the highest support vector machine that showed better classification ability to be our implemented support vector machine.

Table I. Auto Encoder Testing and Training Accuracy

Signal Representing	Average Training Accuracy	Average Testing Accuracy	Classification Algorithm	Simulation Time
Spectrogram	95%	73%	SoftMax Layer	830.93 Seconds
Spectrogram	95.5%	79.14%	ELM (Sigmoid)	379.10 Seconds
Spectrogram	95.5%	82.78%	ELM (ReLU)	379.54 Seconds
Spectrogram	97.237%	83.56%	ELM (RBF)	353.756 Seconds
Spectrogram	89.48%	77.27%	Cubic SVM	1000.409 Seconds
Wavelet	91.42%	45.71%	SoftMax Layer	210.04 Seconds
Wavelet	79.95%	59.88%	ELM (Sigmoid)	265.23 Seconds
Wavelet	83.62%	64.73%	ELM (ReLU)	276.149 Seconds
Wavelet	81.98%	62.70%	ELM (RBF)	257.83 Seconds
Wavelet	70.29%	52.29%	Linear SVM	341.57 Seconds
Wavelet (Scale 5)	98.85%	82.13%	SoftMax Layer	668.355 Seconds
Wavelet (Scale 5)	95.65%	85.416%	ELM (Sigmoid)	402.647 Seconds
Wavelet (Scale 5)	96.98%	86.827%	ELM (ReLU)	276.697 Seconds
Wavelet (Scale 5)	95.59%	85.58%	ELM (RBF)	444.067 Seconds
Wavelet (Scale 5)	96.55%	83.85%	Cubic SVM	495.124 Seconds
Wavelet Packet	98.69%	84.176%	SoftMax Layer	429.52 Seconds
Wavelet Packet	93.3%	86.79%	ELM (Sigmoid)	419.27 Seconds
Wavelet Packet	96.42%	89.41%	ELM (ReLU)	419.27 Seconds
Wavelet Packet	95.59%	87.78%	ELM (RBF)	540.22 Seconds
Wavelet Packet	98.16%	89.707%	Quad SVM	579.713 Seconds

outstanding in our application for all signal representation except for wavelet packet. Both quadratic support vector machine and extreme learning machine, with the rectified linear unit as an activation function, showed a very close performance for wavelet packet signal representation. Extreme learning machine was improved when we replaced sigmoid activation function by Radial basis function and the rectified linear unit. The rectified linear unit activation function for extreme learning machine presented a superior behaviour than radial basis function and sigmoid activation functions for wavelet, Wavelet scale 5 and wavelet packet. However, the rectified linear unit offered better performance than sigmoid and lower accuracy than radial basis function for spectrogram.

Cubic and Quad support vector machine started to result in a good testing accuracy for Wavelet scale 5 and wavelet packet only. Simulation time for support vector machine is relatively longer than other compared classification algorithms. SoftMax layer resulted in a very poor classification for wavelet signal representation, as the testing accuracy was less than 50%. Softmax started to prove its classification ability for Wavelet scale 5 and wavelet packet. Fig.6 shows different P values for different classifiers and Fig.7 shows confidence intervals for each classifier where it was calculated twice. The grey bars were calculated for different classifiers with three signal representation methods (Wavelet Scale5, Wavelet Packet and Spectrogram). While, yellow bars calculated for different classifiers with four signal representation methods (Wavelet, Wavelet Scale5, Wavelet Packet and Spectrogram). The narrowest interval was 2.53% for the extreme learning machine. While the widest one was 6.10% for support vector machine and softmax layer interval reached 5.77%.

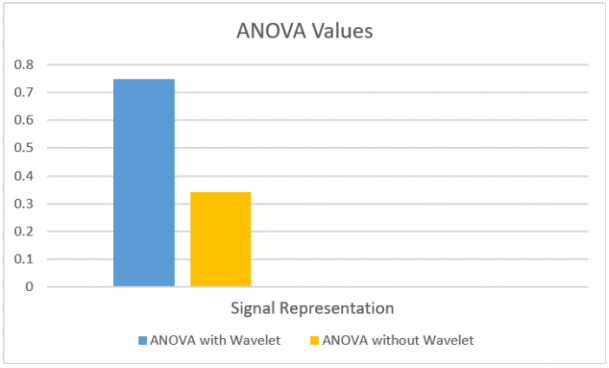


Fig.6. ANOVA values for different Classifiers

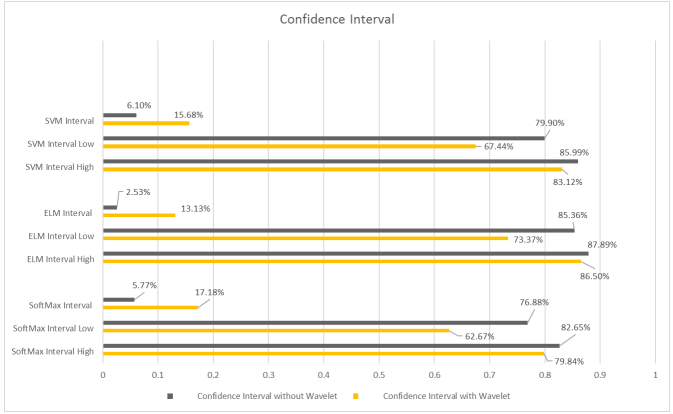


Fig.7. Confidence Intervals for used classifiers

We concatenated a layer of classifier fusion after classification layer. The function of this added layer is to nominate the best-implemented classifier based on the outcomes of accuracies values. This added classifier fusion layer in return enriched our accuracies as displayed in Table II. On the other side, adding classifier fusion layer relatively increased the simulation time than without fusion layer.

Table II: Auto Encoder Classifier Fusion Testing and Training Accuracy

Signal Representing	Average Training Accuracy	Average Testing Accuracy	Classification Algorithm	Simulation Time
Spectrogram	99.53%	91.05%	Classifier Fusion	2118.846 Seconds
Wavelet	96.99%	86.80%	Classifier Fusion	745.67 Seconds
Wavelet (Scale 5)	99.24%	89.02%	Classifier Fusion	1378.41 Seconds
Wavelet Packet	98.70%	92.25%	Classifier Fusion	1656.724 Seconds

7. Conclusion

Sparse autoencoder is just one hidden layer algorithm. Therefore, to establish the concept of deep learning, and take advantage of stacking more than a layer, as the testing set accuracy was less than 50% for one stage only of the autoencoder, we implemented stacked autoencoder that led to verifying deep learning concept and enriching the results accuracy. In addition, applying some signal representation like calculating spectrogram, wavelet and wavelet packet, instead of using raw signal, and introducing the output of these signal representation to the first stage of autoencoder enhanced the performance of the system. Extreme learning machine showed a satisfactory performance on the level of testing accuracy and simulation time. Softmax layer classification resulted in the most mediocre testing accuracy although it consumed longer simulation time than extreme learning machine. Support vector machine produced an excellent testing accuracy but consumed long simulation time. Applying signal representation like Spectrogram, wavelet or wavelet packet improved both training and testing accuracy a lot as both accuracies were much less than 50% when we fed first stage autoencoder by raw data. Multiplying wavelet scale by 5 enhanced the results a lot. As a conclusion, applying any signal representation either in the time domain, frequency domain, or both had a good impact on our training and testing accuracies. We introduced the rectified linear unit as an activation function for extreme learning machine besides already existing functions such as radial basis function and sigmoid. The rectified linear unit was superior in its testing accuracy than both radial basis function and sigmoid one for wavelet, Wavelet scale 5 and wavelet packet signal representation. Moreover, it resulted in a lower testing accuracy than radial basis function but better than sigmoid for spectrogram signal representation.

learnt, by itself, the best features suitable for the application under examination. In addition, since feature extraction and reduction methods varied according to the application so, feature extraction and reduction algorithms were not fixed and needed more experience. In other words, deep learning system should be adaptable to any set of data if the data was accurate and well represented. This brought a new challenge on the scene in regarding representing the data. The data should be represented in a high precision way to expect a good result from implementing deep learning technique.

References

[1] R. Dechter, "Learning While Searching in Constraint-Satisfaction-Problems," presented at the Proceedings of the 5th National Conference on Artificial Intelligence, Philadelphia, 1986.

[2] L. Y. Deng, D, "Deep Learning: Methods and Applications," *Foundations and Trends in Signal Processing,* vol. 7, pp. 3-4, 2014.

[3] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning,* vol. 2, no. 1, pp. 1-127, 2009.

[4] Y. C. Bengio, A.; Vincent, P., "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 8, pp. 1798–1828, 2013.

[5] J. Schmidhuber., "Learning complex, extended sequences using the principle of history compression," *Neural Computation,* vol. 4, pp. 234–242, 1992.

[6] G. E. Hinton, "Deep belief networks," *Scholarpedia,* vol. 4, no. 5, p. 5947, 2009.

[7] A. H. Noriyoshi Uchida, Noboru Sonehara, and Katsunori Shimohara, "EMG pattern recognition by neural networks for multi fingers control," presented at the Engineering in Medicine and Biology Society, 1992 14th Annual International Conference of the IEEE, 1992.

[8] E. M. a. L. M. Alex Andrews, "Optimal Electrode Configurations for Finger Movement Classification using EMG," presented at the 31st Annual International Conference of the IEEE EMBS, 2009.

[9] S. K. Rami N. Khushaba, Dikai Liu, Gamini Dissanayake, "Electromyogram (EMG) Based Fingers Movement Recognition Using Neighborhood Preserving Analysis with QR-Decomposition," presented at the 2011 Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011.

[10] C. A. a. C. C. Gunter R. Kanitz, "Decoding of Individuated Finger Movements Using Surface EMG and Input Optimization Applying a Genetic Algorithm," presented at the 33rd Annual International Conference of the IEEE EMBS, Boston, Massachusetts USA, 2011.

[11] G. B. Ali H. Al-Timemy, Javier Escudero and Nicholas Outram, "Classification of Finger Movements for the Dexterous Hand Prosthesis Control With Surface Electromyography," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 17, no. 03, pp. 608-618, 2013.

[12] W. C. Mochammad Ariyanto, Khusnul A. Mustaqim, Mohamad Irfan, Jonny A. Pakpahan, Joga D. Setiawan and Andri R. Winoto, "Finger Movement Pattern Recognition Method Using Artificial Neural Network Based on Electromyography (EMG) Sensor," presented at the 2015 International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT), Bandung, Indonesia, 2015.

[13] C. W. Daniele Rav`ı, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo and Guang-Zhong Yang, "Deep Learning for Health Informatics," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 4-21, 2016.

[14] T. K. Julius Hannink, Cristian F. Pasluosta, Karl-G ¨unter Gaßmann, Jochen Klucken, and Bjoern M. Eskofier, "Sensor-Based Gait Parameter Extraction With Deep Convolutional Neural Networks," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 85-93, 2016.

[15] T. T. Phuoc Nguyen, Nilmini Wickramasinghe, and Svetha Venkatesh, "Deepr: A Convolutional Net for Medical Records," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 22-30, 2016.

[16] C. W. Daniele Rav`ı, Benny Lo, and Guang-Zhong Yang, "A Deep Learning Approach to on-Node Sensor Data Analytics for Mobile or Wearable Devices," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 56-64, 2016.

[17] Y. Bengio, "Learning Deep Architectures for AI," in "Foundations and Trends in Machine Learning," Universit´e de Montr´eal, Canada2009.

[18] D. J. F. BRUNO A OLSHAUSEN, "Sparse Coding with an Overcomplete Basis Set: Strategy Employed by V1?," *Pergamon,* vol. 37, no. 23, pp. 3311-3325, 1997.

[19] V. a. H. Nair, Geoffrey E, "In Advances in Neural Information Processing Systems," presented at the Advances in Neural Information Processing Systems, 2009.

[20] A. a. N. Coates, Andrew, "The importance of encoding versus training with sparse coding and vector quantization.," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 921–928.

[21] Cheng-Yuan Liou, Jau-Chi Huang, Wen-Chie Yang, "Modeling word perception using the Elman network," *Neurocomputing,* vol. 71, no. 16-18, pp. 3150-3157, 2008.

[22] C.-Y. Liou, Cheng, C.-W., Liou, J.-W., and Liou, D.-R., "Autoencoder for Words," *Neurocomputing,* vol. 139, pp. 84–96, 2014.

[23] G. E. H. a. R. R. Salakhutdinov. (2006) Reducing the Dimensionality of Data with Neural Networks. *SCIENCE*.

[24] F. V. Maryam M Najafabadi, Taghi M Khoshgoftaar, Naeem Seliya, Randall WaldEmail author and Edin Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data,* vol. 2, no. 1, pp. 1-21, 2015.

[25] S. K. a. L. H. Zhou G, "Online incremental feature learning with denoising autoencoders," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 1453–1461.

[26] L. H. Vincent P, Bengio Y, and Manzagol Pierre Antoine, "Extracting and composing robust features with denoising autoencoders," in *25th international conference on Machine learning*, Helsinki, Finland, 2008.

[27] H. L. Pascal Vincent, Isabelle Lajoie, Yoshua Bengio and Pierre-Antoine Manzagol "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *The Journal of Machine Learning Research,* vol. 11, no. 3, pp. 3371-3408 2010.

[28] X. Z. Chen M, Weinberger KQ and Sha F, "Marginalized denoising autoencoders for domain adaptation," in *29th International Conference in Machine Learning*, Edingburgh, Scotland, 2012.

[29] B. A. a. B. Y. Glorot X, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *28th International Conference on Machine Learning* 2011, pp. 513–520.

[30] L. Y. a. F. Weisen, "Application of stacked denoising autoencoder in spamming filtering," *Journal of Computer Applications,* vol. 35, no. 11, pp. 3256-3260, 2015.

[31] I. D. Ervin Sejdic, Jin Jiang "Time-frequency feature representation using energy concentration: An overview of recent advances," *Digital Signal Processing,* vol. 19, no. 1, pp. 153-183, 2009.

[32] R. R. Mergu and S. K. Dixit, "Multi-resolution speech spectrogram," *International Journal of Computer Applications,* vol. 15, no. 4, pp. 28-32, 2011.

[33] S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *The Journal of the Acoustical Society of America,* vol. 119, no. 1, pp. 360-371, 2006.

[34] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *Information Theory,* vol. 36, no. 5, pp. 961 - 1005, 1990.

[35] M. Farye, "WAVELET TRANSFORMS AND THEIR APPLICATIONS TO TURBULENCE," *Fluid Mechanics,* vol. 24, pp. 395–457, 1992.

[36] H. W. a. K.-M. Lau, "Wavelets, Period Doubling, and Time–Frequency Localization with Application to Organization of Convection over the Tropical Western Pacific," *Journal of the Atmospheric Sciences,* vol. 51, no. 17, pp. 2523–2541., 1994.

[37] D. G. a. S. G. H. Philander, "Secular Changes of Annual and Interannual Variability in the Tropics during the Past Century," *Journal of Climate,* vol. 8, pp. 864-876, 1995.

[38] B. W. a. Y. Wang, "Temporal Structure of the Southern Oscillation as Revealed by Waveform and Wavelet Analysis," *Journal of Climate,* vol. 9, pp. 1586-1598, 1996.

[39] N. Gamage, and W. Blumen, "Comparative analysis of lowlevel cold fronts: Wavelet, Fourier, and empirical orthogonal function decompositions," *Monthly Weather Review,* vol. 121, pp. 2867–2878, 1993.

[40] P. F. Sallie Baliunas, Dmitry Sokoloff and Willie Soon, "Time scales and trends in the central England temperature data (1659–1990)," *BrAVELET ANALYSIS OF CENTRAL ENGLAND TEMPERATURE* vol. 24, pp. 1351-1354, 1997.

[41] A. Calderbank, I. Daubechies, W. Sweldens, and B.-L. Yeo, "Wavelet transforms that map integers to integers," *Applied and computational harmonic analysis,* vol. 5, no. 3, pp. 332-369, 1998.
[42] J. Lin and L. Qu, "Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis," *Journal of sound and vibration,* vol. 234, no. 1, pp. 135-148, 2000.
[43] J. Lin and M. Zuo, "Gearbox fault diagnosis using adaptive wavelet filter," *Mechanical systems and signal processing,* vol. 17, no. 6, pp. 1259-1269, 2003.
[44] M. V. Wickerhauser, "INRIA lectures on wavelet packet algorithms," 1991.
[45] R. P. a. G. Wilson, "Application of 'matched' wavelets to identification of metallic transients," in *Proceedings of the IEEE-SP International Symposium Time-Frequency and Time-Scale Analysis*, Victoria, BC, Canada, Canada, 1992: IEEE.
[46] E. S. a. M. Fabio, "The use of the discrete wavelet transform for acoustic emission signal processing," in *Proceedings of the IEEE–SP International Symposium*, Victoria, British Columbia, Canada, 1992: IEEE.
[47] T. P. T. Brotherton, R. Barton, A. Krieger, and L. Marple, "Applications of time frequency and time scale analysis to underwater acoustic transients," in *Proceedings of the IEEE–SP International Symposium*, Victoria, British Columbia, Canada, 1992: IEEE.
[48] M. Y. G. a. D. K. Khanduj, "Time Domain Signal Analysis Using Wavelet Packet Decomposition Approach " *Int. J. Communications, Network and System Sciences,* vol. 3, pp. 321-329, 2010.
[49] J.-G. Han, W.-X. Ren, and Z.-S. Sun, "Wavelet packet based damage identification of beam structures," *International Journal of Solids and Structures,* vol. 42, no. 26, pp. 6610-6627, 2005.
[50] J.-D. Wu and C.-H. Liu, "An expert system for fault diagnosis in internal combustion engines using wavelet packet transform and neural network," *Expert systems with applications,* vol. 36, no. 3, pp. 4278-4286, 2009.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 103-108 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Statistical Approach for Gain Bandwidth Prediction of Phoenix-Cell Based Reflect arrays

Hassan Salti*, Raphael Gillard

Institute of Electronics and Telecommunication of Rennes, INSA de Rennes CS 70839 – 35708 Rennes Cedex 7, France

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
Reflectarray
Bandwidth
Prediction

ABSTRACT

A new statistical approach to predict the gain bandwidth of Phoenix-cell based reflectarrays is proposed. It combines the effects of both main factors that limit the bandwidth of reflectarrays: spatial phase delays and intrinsic bandwidth of radiating cells. As an illustration, the proposed approach is successfully applied to two reflectarrays based on new Phoenix cells.

1. Introduction

This paper is an extension of work originally presented in the 11th European Conference on Antennas and Propagation (EuCAP) where a novel single layer stub-patch Phoenix cell is suggested as a broadband and easy solution to fabricate reflectarray (RA) elements [1].

The Phoenix cell concept was firstly introduced in [2] and many other Phoenix cells were derived later [3-8]. As their name suggests, Phoenix cells are characterized by rebirth capabilities, which means that their geometry comes back to its initial state after a complete 360° phase cycle. This guarantees a smooth evolution of cell geometries over the RA panel and prevents perturbations in the radiation pattern. Phoenix cells are also characterized by a quasi-linear phase response which classifies them as broadband RA cells. Nevertheless, as for other broadband cells, their phase response is still not totally perfect and the proper assessment of the residual phase error versus frequency is still missing in the literature.

In this paper, this last issue is addressed and a new statistical approach for estimating the bandwidth of RA based on Phoenix cells is proposed. The suggested approach relies on the standard deviation of phase errors over the RA panel and combines the effects of both bandwidth limiting factors: the dispersion of spatial phase delays with frequency [9] and the intrinsic limited bandwidth of cells themselves [10]. This standard deviation is shown to provide a promising figure of merit, better than those in [9, 10] where only the maximum phase error due to spatial delays is taken into account.

The paper is organized as follows. In section 2, the new statistical approach is defined. In sections 3 and 4, a bandwidth estimator for RA gain is derived. Finally, in section 5, it is validated by comparison with the simulated bandwidth of a test-case RA based on two different Phoenix cell topologies.

2. Statistical Bandwidth Estimator

Consider an *N*-cell circular RA of diameter *D* with a feeding antenna positioned at distance *F* normally above the array center. The phase of the wave radiated by cell *i* at central frequency f_0 is defined as:

$$\phi_i^{Rad}(f_0) = \phi_i^{Inc}(f_0) + \phi_i^{Sh}(f_0) \quad (1)$$

where $\emptyset_i^{Sh}$ is the phase-shift produced by the cell and $\emptyset_i^{Inc}$ is the phase of the incident wave defined as:

$$\phi_i^{Inc}(f_0) = -2\pi f_0 F\sqrt{1+(\rho_i/F)^2}/c \quad (2)$$

In these equations, ρ_i is the radial distance between the center of the array and cell *i* and *c* is the velocity of light in vacuum.

In order to produce a desired radiation pattern, the required radiated phases $\emptyset_i^{Rad}$ are usually specified at f_0 and the subsequent

*Hassan Salti, Department of Electrical Engineering, Australian College of Kuwait, P.O. Box 1411, Safat 13015, Kuwait, Email: h.salti@ack.edu.kw

phase-shifts $\emptyset_i^{Sh}(f_0)$ are directly deduced from (1) and (2). The cells are selected accordingly and appropriately distributed over the RA.

When the frequency is shifted to $f=f_0+\Delta f$, the phase radiated by cell i is changed by $\Delta\emptyset_i^{Tot}$ as:

$$\Delta\phi_i^{Tot}(f) = \Delta\phi_i^{Inc}(f) + \Delta\phi_i^{Sh}(f) \quad (3)$$

where $\Delta\emptyset_i^{Inc}$ is the predictable deviation of the phase of the incident wave defined as:

$$\Delta\phi_i^{Inc}(f) = -2\pi\Delta f F\sqrt{1+(\rho_i/F)^2}/c \quad (4)$$

while $\Delta\emptyset_i^{Sh}$ is the cell-dependent phase deviation due to the cell's dispersive phase response. At the RA level, these phase errors are responsible for a decrease of the gain at f, and thus for the limited bandwidth.

The bandwidth estimator we propose is derived from the standard deviation of the total phase error. Let $\Delta\emptyset^{Tot}(f)$, $\Delta\emptyset^{Inc}(f)$, and $\Delta\emptyset^{Sh}(f)$ be the statistical variables related to the total, incident and phase-shift errors at f respectively. $\sigma^{Tot}(f)$, $\sigma^{Inc}(f)$ and $\sigma^{Sh}(f)$ are the respective standard deviations. Using (3), the standard deviation of the total error $\sigma^{Tot}(f)$ can be expressed as:

$$\sigma^{Tot}(f) = \sqrt{[\sigma^{Inc}(f)]^2 + [\sigma^{Sh}(f)]^2 + 2\,\mathrm{cov}[\Delta\phi^{Inc}(f), \Delta\phi^{Sh}(f)]} \quad (5)$$

Assuming $\Delta\emptyset^{Inc}(f)$ and $\Delta\emptyset^{Sh}(f)$ are uncorrelated, which is the case when the synthesis process is done at f_0 only, as usually applied in the literature [11-15], the covariance term reduces to zero and $\sigma^{Tot}(f)$ reduces consequently to:

$$\sigma^{Tot}(f) = \sqrt{[\sigma^{Inc}(f)]^2 + [\sigma^{Sh}(f)]^2} \quad (6)$$

3. Bandwidth Estimator: Incident Phase Errors

3.1. Standard Deviation of Incident Phase Errors

Defining:

$$S_i = \sqrt{1+(\rho_i/F)^2} \quad (7)$$

(4) can be reformulated as:

$$\Delta\phi_i^{Inc}(f) = -2\pi\Delta f F S_i/c \quad (8)$$

Due to the mathematical properties of standard deviation, $\sigma^{Inc}(f)$ can be derived as:

$$\sigma^{Inc}(f) = 2\pi|\Delta f|F\sigma_S/c \quad (9)$$

where: $$\sigma_S = \sqrt{E[S^2] - (E[S])^2} \quad (10)$$

E[S] and *E[S²]* are the first statistical moments and can be calculated as:

$$\begin{aligned} E[S] &= (2/D)\int_0^{D/2}\sqrt{1+(\rho/F)^2}\,d\rho \\ &= 0.5\sqrt{1+(D/2F)^2} \\ &\quad + (F/D)\ln[(D/2F)+\sqrt{1+(D/2F)^2}] \end{aligned} \quad (11)$$

$$\begin{aligned} E[S^2] &= (2/D)\int_0^{D/2}[1+(\rho/F)^2]\,d\rho \\ &= 1+0.33(D/2F)^2 \end{aligned} \quad (12)$$

Note that σ_S does not depend on frequency but only on the dimensions of the RA, which is consistent with other criteria in the literature [10]. Note also that, in (11) and (12), a rectangular lattice is considered and ρ_i is consequently supposed to vary uniformly in the [0, D/2] range. In addition, though ρ_i is normally a discrete variable, it is assumed here to vary continuously. This assumption is reasonable since the inter-element spacing is usually a small fraction of λ_0 (which is much lower than D/2). Furthermore, as in [9] and [10] and for simplicity reasons, S_i and the resulting σ_S are calculated for a centered fed RA. Different expressions could easily be established for offset configurations.

3.2. Bandwidth Estimator

We now investigate how the gain deteriorates with respect to σ^{Inc}. To do so, we consider the gain at broadside for a test-case RA with 12mm spacing (i.e. $0.5\lambda_0$ at the center of the [11-14] GHz frequency band). Edge tapering is supposed to be -12dB and different antenna configurations are considered: F/D=0.6 and 0.8 with D varying from 0.28m≈ $12\lambda_0$ to 1m ≈ $42\lambda_0$. For each couple (F/D; D), the reflected field (phase and magnitude) in the aperture is calculated and the associated gain is derived using simple array theory (as in [16]). Simultaneously, $\sigma^{Inc}(f)$ is also computed as the standard deviation of all phase errors. Finally, Figure 1 gives the representation of the normalized simulated gain $G(f)/G(f_0)$ versus the corresponding standard deviation σ^{Inc}.

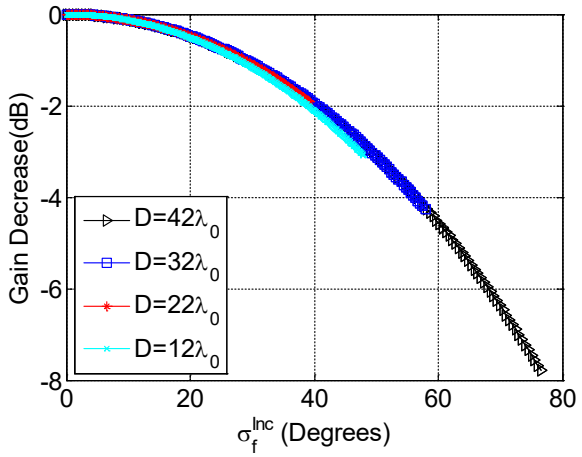


Figure 1. Simulated gain decrease with respect to the standard deviation of incident phase errors (F/D=0.8 or F/D=0.6).

An important conclusion from Figure 1 is that all curves superimpose whatever the particular values of *D, F* or *f*. This demonstrates that σ^{Inc} is a reliable estimator for the bandwidth since it directly reflects the gain decrease. Figure 1 also shows that a 1dB gain-drop approximately corresponds to σ^{Inc}=π/6. As a consequence, the upper frequency f_{max} of the -1dB bandwidth may be derived simply by replacing $\sigma^{Inc}(f)$ by π/6 in (9), leading to:

$$B_{Inc} = 2(f_{\max} - f_0)/f_0 = c/(6Ff_0\sigma_S) \qquad (13)$$

where

$$f_{\max} = f_0 + c/(12F\sigma_S) \qquad (14)$$

Equation (13) will thus be used as a bandwidth estimator. At this stage, it does not depend on any particular cell topology but only on the spatial phase delay error. As will be seen now, this initial estimator can advantageously be replaced by a more sophisticated one that also accounts for the phase dispersion of the used RA cells. In what follows, the case of cells with an ideal phase response is considered.

4. Bandwidth Estimator: Total Phase Errors

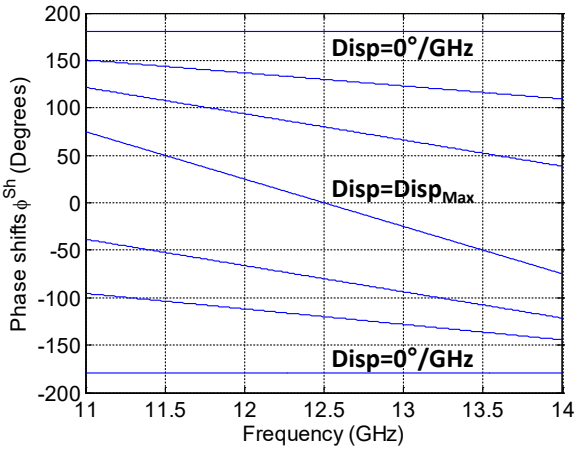


Figure 2. Linear phase response of an ideal Phoenix cell.

The ideal cell we suggest here is quite representative of Phoenix cells as will be shown in section 5. It is supposed to provide a phase range of at least 360° at f_0 and perfect linear variations with respect to frequency. Figure 2 shows the phase response versus frequency of the ideal cell. As such, the phase-shift error for cell *i* at frequency *f* can be expressed as:

$$\Delta\phi_i^{Sh}(f) = \phi_i^{Sh}(f) - \phi_i^{Sh}(f_0) = Disp_i.\Delta f \qquad (15)$$

where $Disp_i$ is the dispersion that is supposed to vary uniformly in the interval [0-$Disp_{max}$]. More specifically, $Disp_i$ is considered to be equal to 0°/GHz for phase-shifts $\emptyset_i^{Sh}(f_0)$= ±180° and to reach its maximum (i.e. $Disp_{max}$) when $\emptyset_i^{Sh}(f_0)$= 0°. Note that the perfect Phoenix cell is obtained when $Disp_{max}$= 0°/GHz as all its phase states would be perfectly parallel.

Assuming all phase-shifts are equally probable on the radiating aperture, $\sigma^{Sh}(f)$ can be expressed as:

$$\sigma_f^{Sh} = |\Delta f| Disp_{\max}/\sqrt{12} \qquad (16)$$

Using (9) and (16) in (6), a generalized bandwidth estimator $\sigma^{Tot}(f)$ accounting for both types of errors is derived:

$$\sigma_f^{Tot} = |\Delta f|\sqrt{(2\pi F\sigma_S/c)^2 + (Disp_{\max}/\sqrt{12})^2} \qquad (17)$$

Finally, $\sigma^{Tot}(f)$ is set to π/6 as already done in Part 3.2 to obtain the generalized bandwidth estimator:

$$B_{Tot} = B_{Inc}/\sqrt{1 + 0.75(Disp_{\max}.f_0B_{Inc}/180°)^2} \qquad (18)$$

To validate (18), the gain decrease simulation described in Section 3.2 is repeated (as in [16]), now accounting for both types of phase errors. The associated simulated -1dB gain bandwidth is then extracted and compared to the theoretical value predicted from (18). In this study, *F/D* is set to 0.8, *D* varies from 0.16m to 1m (i.e. ~7λ_0 to 42λ_0) and $Disp_{max}$ varies from 0°/GHz to 100°/GHz. Figure 3 shows that the difference between simulation and theory is less than 3%, even for the highest dispersion and the smallest diameter values.

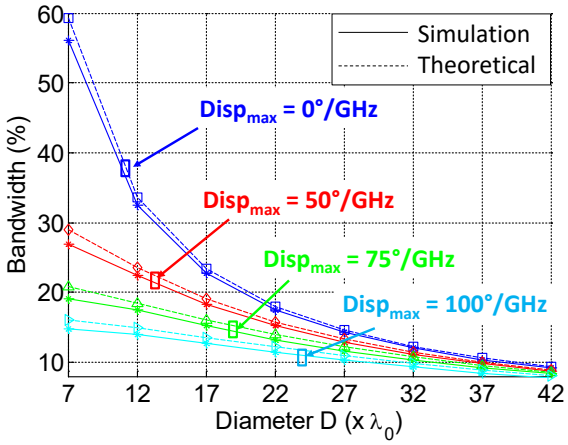


Figure 3. Total phase errors effects: simulated and theoretical bandwidths (F/D=0.8).

As a conclusion, (18) appears to be a reliable bandwidth estimator. In practice, it can be used to define the maximum acceptable cell dispersion for a given application. As an example, for a RA with *D*=22λ_0 and *F/D* = 0.8, the cell dispersion should be less than 50°/GHz to ensure a 15% bandwidth.

5. Practical Validation

In practice, the phase response of Phoenix cells is not purely linear as in the previously-used ideal cell model. Therefore, to assess the validity of our approach, the actual performance of two recently-proposed Phoenix cells is assessed and compared to those obtained using (18).

The two cells are designed to operate around a central frequency f_0 = 12.5GHz, with $\lambda_0/2$ spacing at f_0. Both cells are printed on a Duroïd substrate with 2.17 dielectric constant and backed by a ground plane. The substrate height is fixed to 4 mm.

This corresponds to approximately $\lambda_0/(4\sqrt{\varepsilon_r})$, which means the reflected phase is close to 0° if the cell is transparent.

To extract the phase responses, both cells are simulated using ANSYS-HFSS software assuming normal incidence and local periodicity. As Phoenix cells allow for smooth evolution of cell geometries over the RA panel, it is assumed that the phase responses obtained by simulation are valid for finite reflectarray configurations [2].

5.1. Cell 1: Slot – Patch Phoenix Cell

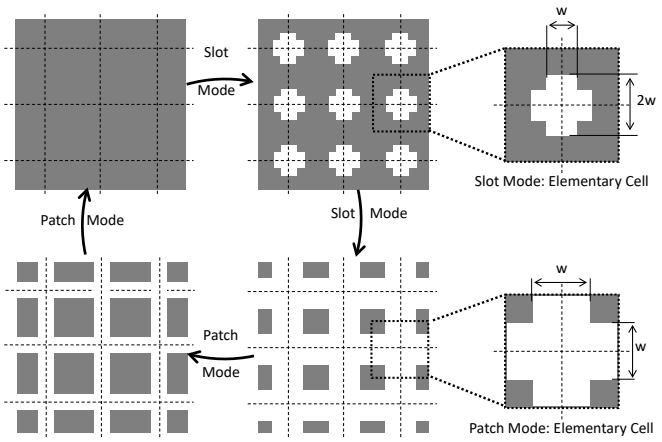


Figure 4. Slot-Patch Phoenix Cell: Rebirth cycle

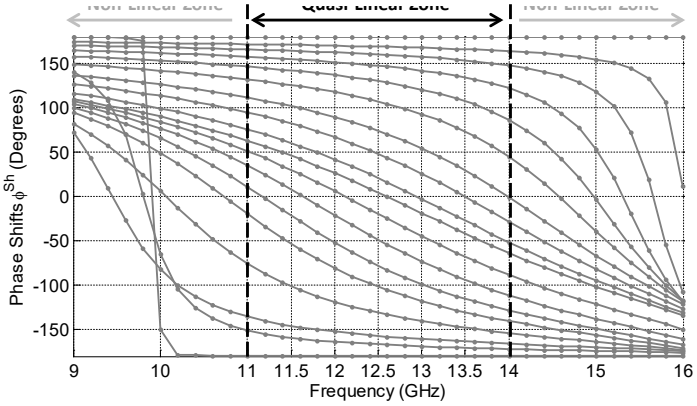


Figure 5. Slot-Patch Phoenix Cell: Phase response.

The first cell, the Slot – Patch Phoenix cell was initially introduced in [8] and its performance was improved in [7]. The cell cycle is illustrated in Figure 4. The initial cell consists of a ground plane providing a 180° phase shift, whatever the frequency. The first mode of operation, or slot mode, is obtained by opening a crossed-shaped slot with variable length and width in the ground plane. For simplicity reasons, the length of the cross is fixed as twice its width. The slot mode ends when the slot arms reach the borders of the cell, thus defining square patches. The operating mode then switches to a patch mode. In this second mode, the length of the pre-opened slot is frozen and only the width of the slot w is decreased. The patch mode ends when the slot vanishes, taking back the cell to its initial geometry and opening the door for a new cycle.

The obtained phase response is presented in Figure 5. At the central frequency, the slot mode provides phase shifts between 0° and 180° while the patch mode completes the remaining phase range between 0° and -180°. The use of complementary modes provides a phase response that is quasi-linear within a 24% bandwidth around 12.5GHz. The maximum dispersion is 53°/GHz. This phase response fits well with the ideal model used to derive (18), although the linearity is not perfect.

5.2. Cell 2: Stub – Patch Phoenix Cell

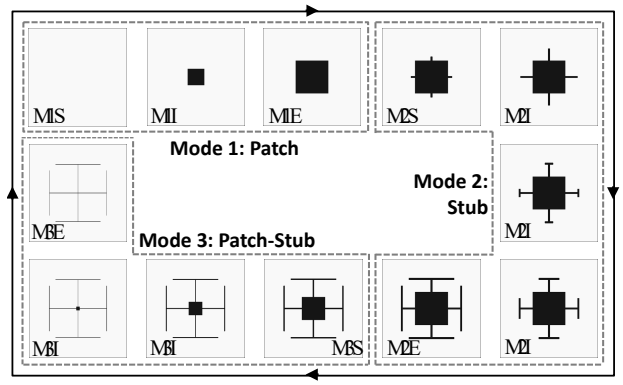


Figure 6. Patch-Stub Phoenix cell: Rebirth cycle (MαS: Mode α's Start; MαI: Mode α's Intermediate states; MαE: Mode α's End).

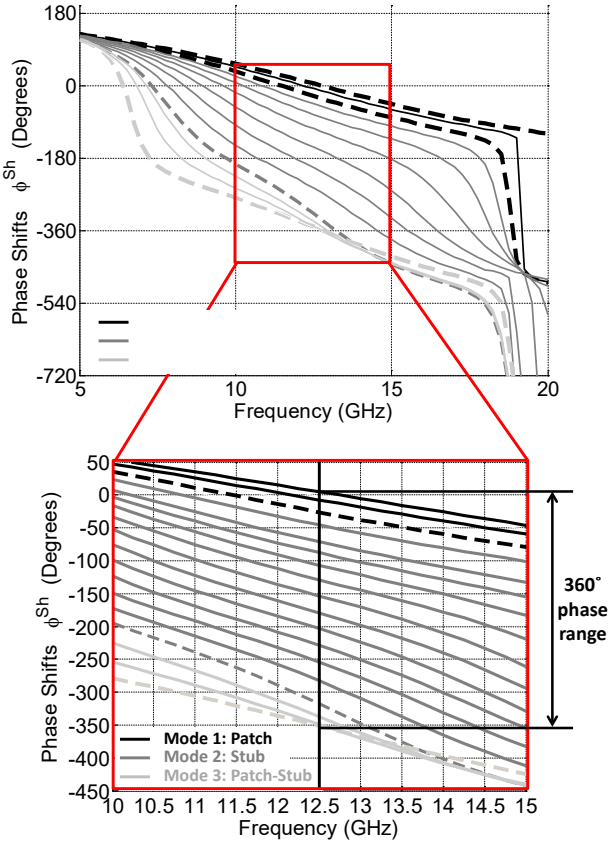


Figure 7. Patch-Stub Phoenix cell: Phase response.

The second cell, the Stub-Patch Phoenix cell was introduced recently in [1]. It improves the bandwidth further due to the three possible operating modes it offers: a patch mode, a stub mode, and a combined patch-stub mode.

As illustrated in Figure 6, the cycle starts with a pure Duroïd substrate backed by a ground plane providing an initial phase shift close to 0° at f_0. The first mode of operation, namely the patch mode, is obtained by inserting a square patch at the center of the cell (cf. Figure 6 – Mode 1). The phase shift produced by the cell is controlled by increasing the patch size until it reaches a maximum predefined value.

In mode 2, namely the stub mode, the patch size is frozen to this maximum value and four open-circuited stubs are grown perpendicularly to the patch from the center of its edges (cf. Figure 6 – Mode 2). A T-shape is used for the stubs when the total metal length approaches the limit allowable by the inter-element spacing. In this mode, the phase shift is thus controlled by the length of the stubs.

The mode then switches to mode 3, namely the patch-stub mode, during which the stub-loaded patch shrinks gradually until both patch and stubs disappear completely allowing the cell to rebirth and to start a new cycle (cf. Figure 6 – Mode 3). In this mode, the phase shift is controlled by the shrinking ratio.

The phase response of the suggested cell in all modes is summarized in Figure 7. Dashed curves represent the start/end of a mode (i.e. MαS/MαE) and continuous lines represent intermediate states (i.e. MαI). As can be noticed, a phase range of 360˚ is achieved at f_0. Within a frequency band ranging from 10 to 15GHz (40%), the phase response is almost linear. The maximum phase dispersion is obtained at the transition from mode 2 to mode 3 and is equal to 55˚/GHz. The minimum dispersion is obtained at the beginning of mode 1 and is equal to 21˚/GHz. Compared to cell 1, cell 2 exhibits a better linearity of phase response. On the other hand, its minimum dispersion is not zero as required by the previously-used ideal cell model. However, this model still applies if we replace the maximum dispersion by the relative maximum dispersion, defined as the difference between the maximum and minimum dispersions. For cell 2, it is then equal to 34˚/GHz.

5.3. Performance Evaluation and Comparison

The previously-described Phoenix cells are now consecutively used as the radiating element in our test-case RA (F/D=0.8, variable D). The bandwidth is calculated by simulations as in [16] and compared to that given by (18). For this theoretical study, the maximum dispersion $Disp_{max}$ in (18) is set to 53°/GHz for cell 1 and 34°/GHz for cell 2.

Figure 8 summarizes all theoretical and simulation results. The results show a remarkable agreement between simulation and theoretical curves for a given cell. The slight discrepancy is mainly due to the linearity assumption in the ideal model which is not fully respected by realistic Phoenix cells. It is less than 5% for cell 1 and 3% for cell 2. As expected, a smaller error is obtained for cell 2 as it offers a better linearity of phase response.

Formula (18) is hence a reliable bandwidth estimator, even for realistic phase-shifting cells, provided that they are characterized by a quasi-linear response. As a consequence, it can be advantageously used to define the maximum dispersion allowed for a Phoenix cell to comply with given bandwidth specifications or to predict a Phoenix cell's performance in a RA configuration.

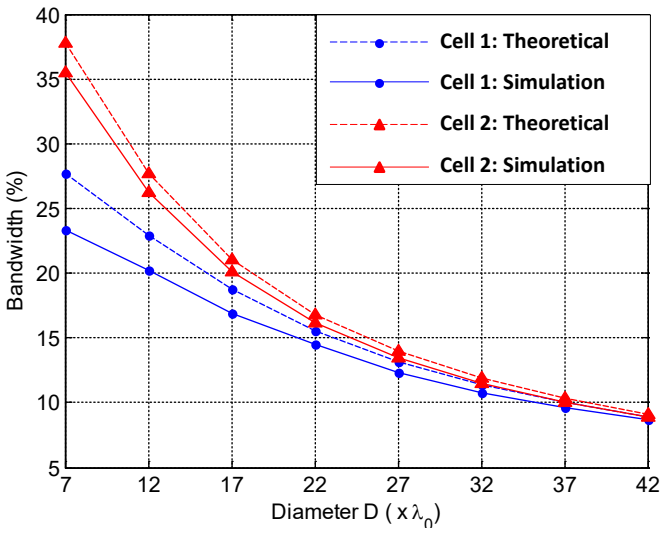


Figure 8. A Realistic study: simulated versus theoretical bandwidth of realistic Phoenix cells (F/D=0.8)

6. Conclusion

Bandwidth limitation of RA results from both the effect of various path delays between cells and source on one side, and intrinsic narrow bandwidth of cells themselves on the other side. The first phenomenon had been significantly investigated in the literature. For the second one, the usual solution relies on the use of broadband cells providing linear and parallel phase states. As this ideal characteristic is never met perfectly, this paper has defined a new approach to assess the effect of imperfection in the phase response of broadband RA cells.

We firstly proposed the standard deviation of phase errors over the array as an efficient criterion to predict RA's bandwidth when accounting for both its limiting phenomena. This criterion has then been formulated and validated for an ideal cell model with linear but non-parallel phase states. Finally, it has been successfully applied to realistic and novel Phoenix-cells. The suggested approach has thus been demonstrated as a powerful tool to help the designer in the selection of appropriate cells before entering the complex RA optimization process.

References

[1] H. Salti, R. Gillard, "A Single Layer Stub-Patch Phoenix Cell for Large Band Reflectarrays" in 11th European Conference on Antennas and Propagation, Paris, France, 2017. https://doi.org/10.1109/LAWP.2011.2108633

[2] L. Moustafa, R. Gillard, F. Peris, R. Loison, H. Legay, E. Girard, "The Phoenix Cell: A New Reflectarray Cell With Large Bandwidth and Rebirth Capabilities" IEEE Antennas and Wireless Propagation Letters, **10**, 71-74, 2011. https://doi.org/10.1109/LAWP.2011.2108633

[3] R. Deng, F. Xu, F. Yang, M. Li, "Single-Layer Dual-Band Reflectarray Antennas With Wide Frequency Ratios and High Aperture Efficiencies Using Phoenix Elements" IEEE Transactions on Antennas and Propagation, **65**(2), 612-622, 2017. https://doi.org/10.1109/TAP.2016.2639023

[4] R. Deng, S. Xu, F. Yang, M. Li, "Design of a Low-Cost Single-Layer X/Ku Dual-Band Metal-Only Reflectarray Antenna" IEEE Antennas and Wireless Propagation Letters, **16**, 2106 - 2109, 2017. https://doi.org/10.1109/LAWP.2017.2698099

[5] Q. Wang, Z. Hai Shao, Y. J. Cheng, P. K. Li, "Ka/W Dual-Band Reflectarray Antenna for Dual Linear Polarization" IEEE Antennas and Wireless Propagation Letters, **16**, 1301 - 1304, 2017. https://doi.org/10.1109/LAWP.2016.2633289

[6] R. Deng, F. Yang, S. Xu, and M. Li, "A low-cost metal-only reflectarray using modified slot-type Phoenix element with 360˚ phase coverage" IEEE Transactions on Antennas and Propagation, **64**(4), 1556-1560, 2016. https://doi.org/10.1109/TAP.2016.2526258

[7] H. Salti, R. Gillard, "Slot-Patch Cell with Low Phase Distortion for Large Band Reflectarrays" in Proc. of the Fifth International Conference on Digital Information and Communication Technology and its Applications, Beirut, Lebanon, 2015. https://doi.org/10.1109/DICTAP.2015.7113196

[8] T. Makdissy, R. Gillard, E. Fourn, E. Girard, H. Legay, "A patch-slot combination approach for large band reflectarrays" in IEEE European Microwave Conference, Amsterdam, Netherlands, 2012. https://doi.org/10.23919/EuMC.2012.6459392

[9] M. E. Bialkowski, K. H. Sayidmarie, "Bandwidth considerations for a microstrip reflectarray" Progress In Electromagnetics Research B, **3**, 173–187, 2008. http://dx.doi.org/10.2528/PIERB07120405

[10] D. M. Pozar, "Bandwidth of reflectarrays" Electronics Letters, **39**(21), 1490-1491, 2003. https://doi.org/10.1049/el:20030990

[11] D. Cadoret, A. Laisné, R. Gillard, and H. Legay, "A new reflectarray cell using microstrip patches loaded with slots" Microwave and Optical Technology Letters, **44**(3), 270-272, 2005. http://onlinelibrary.wiley.com/enhanced/exportCitation/doi/10.1002/mop.20608

[12] S. D. Targonski and D. M. Pozar, "Analysis and design of a microstrip Reflectarray using patches of variable size" in IEEE Symposium on Antennas and Propagation. Seattle, WA, USA, 1994. https://doi.org/10.1109/APS.1994.408184

[13] H. Salti, R. Gillard, R. Loison, and L. Le Coq, "A Reflectarray Antenna Based on Multiscale Phase-Shifting Cell Concept," IEEE Antennas and Wireless Propagation Letters, **8**, 363-366, 2009. https://doi.org/10.1109/LAWP.2008.2006073

[14] M. Bozzi, S. Germani, and L. Perregrini, "Performance comparison of different element shapes used in printed reflectarrays," IEEE Antennas and Wireless Propagation Letters, **2**, 219–222, 2003. https://doi.org/10.1109/LAWP.2003.819687

[15] Kai Zhang, Yangyu Fan, Jiadong Xu, Chen Qu, "Design of broadband, low cost single layer reflectarray using phoenix cell," in TENCON conference, Xi'an, China, 2013. https://doi.org/10.1109/TENCON.2013.6718981

[16] H. Salti, E. Fourn, R. Gillard, "Minimization of MEMS breakdowns effects on the radiation of a MEMS based reconfigurable reflectarray," IEEE Transactions on Antennas and Propagation, **58**(7), 2010. https://doi.org/10.1109/TAP.2010.2048861

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 109-114 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Impact of Crosstalk on Signal Integrity of TSVs in 3D Integrated Circuits

Shadi MS. Harb*[1], William R. Eisenstadt[2]

[1]*Intel Corporation, Hillsboro, OR 97124, USA*

[2]*University of Florida, Department of Electrical & Computer Engineering, Gainesville, FL 32611, USA*

ARTICLE INFO

Article history:
Received: 02 November, 2017
Accepted: 24 December, 2017
Online: 30 January, 2018

Keywords:
3D Interconnect
Crosstalk
Cross-coupled
TSVs
3D Test
Electrical Characterization

ABSTRACT

Through-Silicon-Vias (TSVs) are utilized for high density 3D integration, which induce crosstalk problems and impact signal integrity. This paper focuses on TSV crosstalk characterization in 3D integrated circuits, where several TSV physical and environmental configurations are investigated. In particular, this work shows a detailed study on the influence of signal-ground TSV locations, distances and their structural configurations on crosstalk. Embedded 3D testing circuits are also presented to evaluate the coupling effects between adjacent TSVs such as crosstalk induced delay and glitches for different crosstalk modes. Additionally, A 3D parallel Ring Oscillators testing structure is proposed to provide crosstalk strength coupling indicator between adjacent TSVs. Simulation results are conducted using a 3D electromagnetic field solver (HFSS) from Ansoft Corporation and a Spice-like simulator (ADS) from Keysight Technologies Corporation based on MIT 0.15µm 3DFDSOI process technology.

1. Introduction

3D Interconnect is the promising technology [1]-[4], which includes Through-Silicon-Vias (TSVs) to connect vertically stacked semiconductor chips with shortest paths, which means lowest inductance and conduction loss, to both signals and power supplies. In spite of these benefits, the signal integrity issues in TSVs become the major challenges in 3D designs [5-6]. The goal of TSVs or 3D Vias development is to acquire high chip density. Therefore, the density of the 3D Vias is also high. In this environment, a crosstalk problem appears between two adjacent signal 3D Vias (Aggressor and Victim). Studies show that the coupling problem is not negligible in TSVs because of the relatively large diameter and small pitch, which results in non-negligible TSV-to-TSV coupling that degrades significantly the 3D circuit performance. Hence, it becomes very essential to precisely model and evaluate the electrical characteristics of TSVs [7]-[9] to analyze signal integrity (SI), and crosstalk of adjacent TSVs under the conditions of various structures and configurations.

*Corresponding Author: Shadi MS. Harb, Intel Corporation, Hillsboro, OR 97124, USA | Email: shadiharb@ieee.org

In this paper, the electrical characteristics of 3D interconnect, based on our previous work [10], is presented to characterize signal integrity effects of 3D crosstalk for different TSVs placement and configurations. 3D Vias based on 0.15 µm 3DFDSOI process from MIT Lincoln lab [11] are used as a device under test (DUT) for crosstalk characterization where a 3D full wave simulator such as HFSS from Ansoft Corporation is used to extract and predict the electrical characteristics of TSVs in the frequency domain (S-Parameters) and a Spice-like simulator such as ADS from KEYSIGHT technologies to evaluate the TSVs transient response in the time domain (Eye-diagram). Additionally, embedded 3D testing applications are proposed to characterize the TSV's signal integrity effects and the impact of TSVs on the 3D circuit performance after fabrication. A 3D circuit test is presented to evaluate the coupling effects between adjacent TSVs such as induced-delay and glitches [12-13] for different crosstalk modes. Additionally, a consecutive triggered parallel Ring Oscillators (ROs) testing structure is proposed to provide a crosstalk coupling indicator between adjacent TSVs.

The paper is organized as follows: section 2 discusses the 3D Full wave modeling for TSV and the simulation setup. A detailed study of crosstalk for different physical and environmental TSVs configurations is given in Section 3. 3D Crosstalk embedded

testing applications are given in Section 4. Section 5 concludes the paper.

2. A 3D Full Wave Modeling for TSVs

In order to evaluate the electrical characteristics of a TSV depending on structural parameters such Via pitch, Via height, and Via size, the 3D interconnect based on MIT 0.15µm 3DFDSOI technology is used as DUT to model and characterize crosstalk in different testing configurations. The vertical connection in this technology is slightly different from the standard Through-Silicon-Via (TSV), which is a square shape via made from Tungsten material and fully surrounded by oxide; thus, it is simply called a 3D Via. The 3D Via pitch is around 3.325 µm (distance between the centers of two 3D Vias), 7.34 µm TSV height, 1.25 µm x 1.25 µm TSV size. The physical size of 3D Via after fabrication is estimated to be around 2 µm for the top dimension and 1 µm for the bottom dimension. The 3D Via was simulated using a 3D full wave simulator (HFSS from Ansoft Corporation), which generates the S-Parameters of the structural model of the Via and a Spice type simulator (ADS from Keysight Corporation) to predict the electrical characteristics of 3D Vias in the time domain (Eye-diagram). Figure 1 presents a pair of TSVs structure using the HFSS simulator.

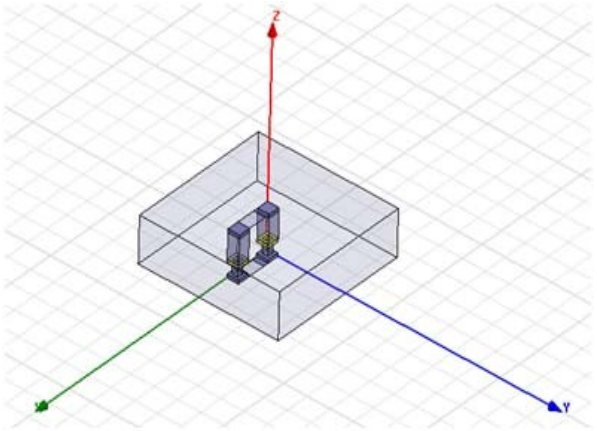


Figure 1. Physical Structure of a pair of TSVs using 3D full wave simulator

Usually, an interconnection line is characterized using S-Parameters. S11 and S21 are the S-parameters reflection and transmission coefficients respectively, which are typical characteristics of an interconnection. The evaluated S11 and S21 magnitudes for the electrical characteristics of a 3D Via based on the default parameters are shown in Figure 2.

The transmitted data stream through the 3D Via was simulated with the evaluated S-parameters from the 3D full wave simulator (HFSS). The Eye-diagram of the transmitted data stream was evaluated for 10^7-1 pseudo random bits sequence (PRBS) using the Spice type circuit simulator (ADS). The source for the simulation has 1.5 Vp-p and 50Ω source termination. The 3D Via is terminated by a shunt connected 50 Ω resistor and 1pF capacitor. The Eye-diagrams of 2 Gbps, and 10 Gbps PRBSs are shown in Figure 3. Also, all PRBSs were assumed that they have 10% rising and falling times

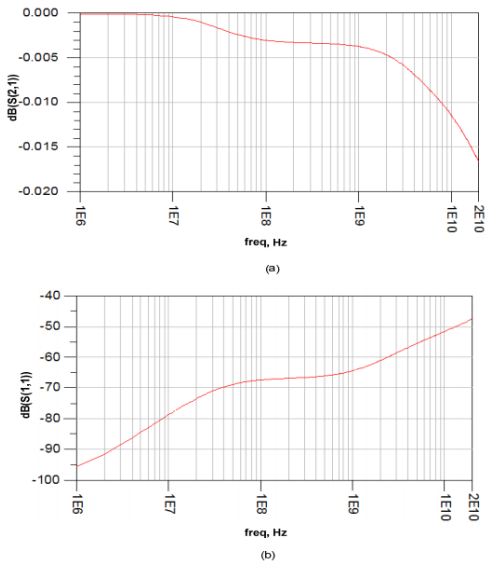


Figure 2. Evaluated S21 magnitude (a) and S11 magnitude (b) of 3D via using HFSS

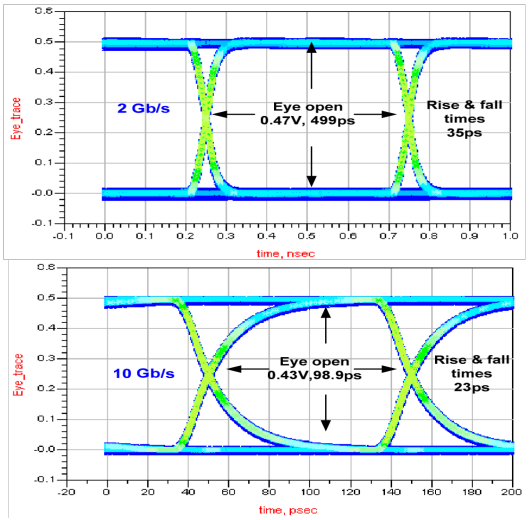


Figure 3. The simulated eye-diagrams of 2Gbps (a), and 10 Gbps (b) PRBSs after passing through the 3D Via

3. A Crosstalk between 3D Vias

3.1. The Influence of 3D Via Locations, and Distances on Crosstalk

Crosstalk is evaluated depending on the distance of two signal Vias and the location of two GND Vias as shown in Figure 4.

Five distances (4 µm, 8 µm, 16 µm, and 32 µm) based on SGSG configuration between the two signal Vias have been simulated. As expected, if the distance between the two signal 3D Vias is larger, the crosstalk level is going down as shown in Figure 5. The effect of the distance of GND Vias with respect to the signal via on the crosstalk is also evaluated as shown in Figure 6 with four distances: 4 µm, 8 µm, 16 µm, and 32 µm. The results show an increase in the magnitude of the crosstalk as the distance of the reference via (GND) increases. Also it is shown in Figure 7 that the crosstalk magnitude of SGGS (i.e. the cross locations of

the signal and ground Vias) configuration is smaller than that of SGSG configuration comparing two same distance cases. The difference of the two cases is almost 10dB. This is a very interesting point for 3D designers to keep in mind, because just changing the 3D Via role can reduce the crosstalk magnitude especially in high frequency applications, where crosstalk problem is very critical to obtain the maximum system performance.

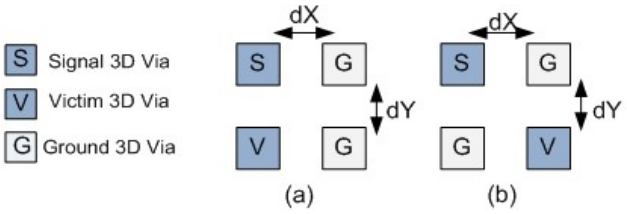


Figure 4. Location of signal Vias and GND Vias for crosstalk evaluation, (a) SGSG, (b) SGGS configurations

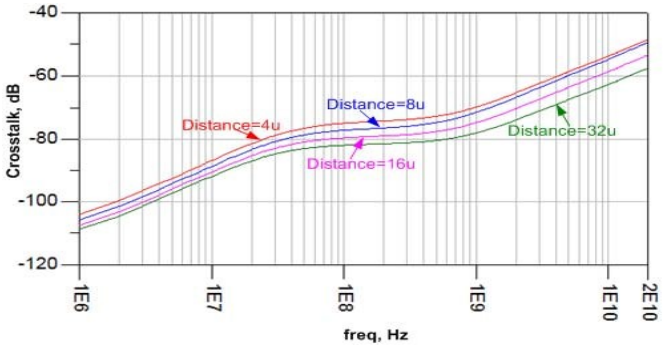


Figure 5. Comparison of five locations of 3D Vias. The distances 4 µm, 8 µm, 16 µm, and 32 µm between two signal Vias are associated to the graphs from top to bottom respectively

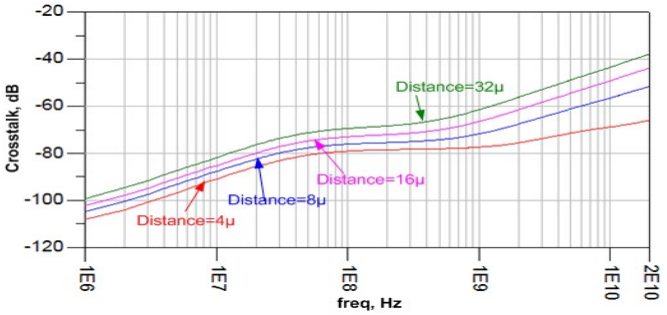


Figure 6. Comparison of three locations of GND Vias. The distances 4 µm, 8 µm, 16 µm, and 32 µm between two GND Vias are associated to the graphs from bottom to top respectively

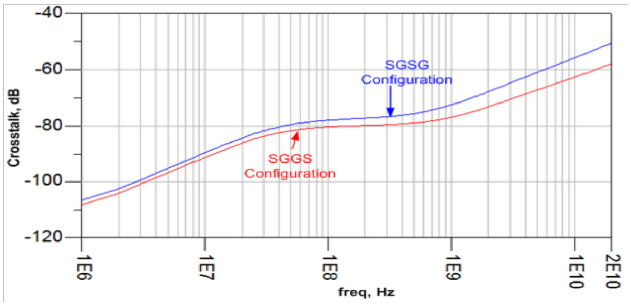


Figure 7. Comparison between SGSG and SGGS configurations for 10 µm distance, SGSG and SGGS are the top and bottom graphs respectively

3.2. 3D Via Crosstalk in Structures with Different Configurations

The geometry of the return current path may be one of the most efficient methods to affect the crosstalk between a signal and victim Via. Four different configurations have been investigated (1) Two Via pairs in a straight line (Figure 8(a)), (2) Two Via pairs placed opposite to each other (Figure 8(b)), (3) A signal Via with two reference via placed opposite to a victim Via with two ground Vias (Figure 8(c)), (4) A signal and victim via, each with three reference Vias as a return current path (Figure 8(d)).

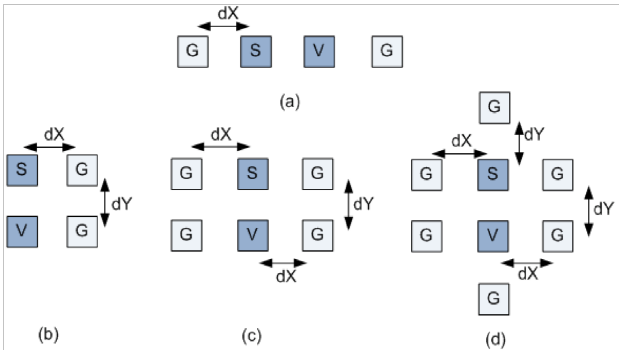


Figure 8. Comparison of three different configurations, (a) Configuration #1, (b) Configuration #2, (c) Configuration #3, and (d) Configuration #4

In configuration 2, the signal Via has a slightly lower inductance than it would in configuration 1 because the second Via is close enough to the signal Via to have a slight impact on its inductance. Configuration 4 has the lowest inductance because it has the most well defined return current path. This lowering of the inductance will also lower the near-end crosstalk as shown in Figure 9. Also it is shown that the highest crosstalk as predicted, comes from configuration 1. Only slightly lower is the crosstalk from configuration 2. Then, there is a significant decrease in crosstalk when the extra two reference Vias are added in configuration 3, and a slight decrease further when the third reference Via is added to the victim and signal Vias in configuration 4. The reduction in crosstalk from adding additional reference Vias is almost 4 dB.

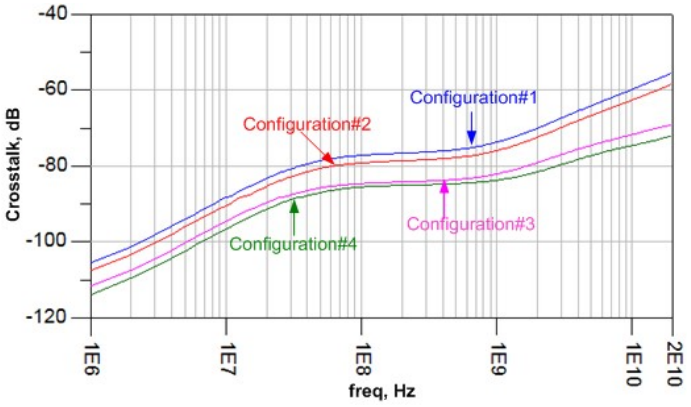


Figure 9. Crosstalk between four different configurations

Until this point, only isolated 3D Vias with respect to ground Vias have been considered. Potential coupling in the 3D Vias in the discontinuity between the 3D via and the transmission line may be an important effect to consider. In Figure 10, a new Ansoft HFSS model is presented that accounts for the discontinuity between the 3D via and the transmission line.

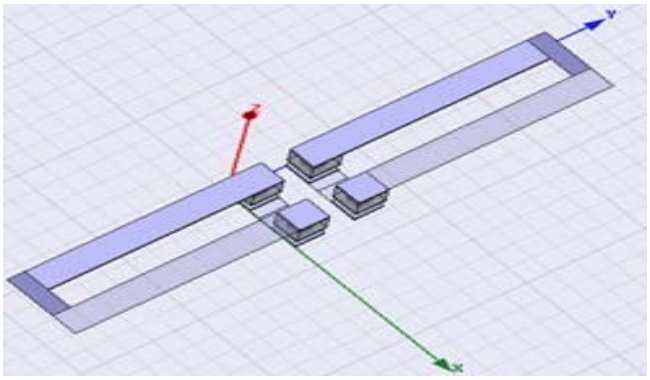


Figure 10. A TSV coupling model that includes the transmission lines

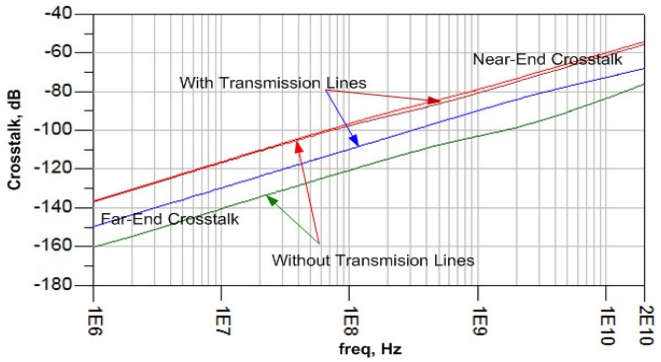


Figure 11. Comparison of crosstalk between TSVs with and without the inclusion of transmission lines

Figure 11 shows that, according to the full wave simulations, there is no measurable difference between the near-end crosstalk as a result of the discontinuity between the transmission lines and the 3D Vias. The difference between the two simulations is very small over the entire frequency range. However, the discontinuity between the transmission lines and 3D Vias will increase the far-end crosstalk. On the other hand, the far-end crosstalk is more minimal than the near-end crosstalk, less significant and never exceeds -50 dB as shown in the simulation.

4. 3D Embedded Crosstalk Test Applications

4.1. 3D Testing Circuit for Crosstalk Induced-Delay and Glitches

In this embedded test application, the coupling effects between adjacent TSVs such as induced-delay and glitches can be investigated for different crosstalk modes. As shown in Figure 12, high speed signals can be fired through three adjacent TSVs at each tier using a multi-edge delay generator circuit. A Mux and Tristate circuits are used to control which signals are active from which tier. The complementary signals are also generated from the delay generator to cover different crosstalk modes. In order to study the effect of phase shifting the aggressor signal on crosstalk induced-delay cancellation, the multi-edge delay generator is used to fine control the delay between adjacent signals.

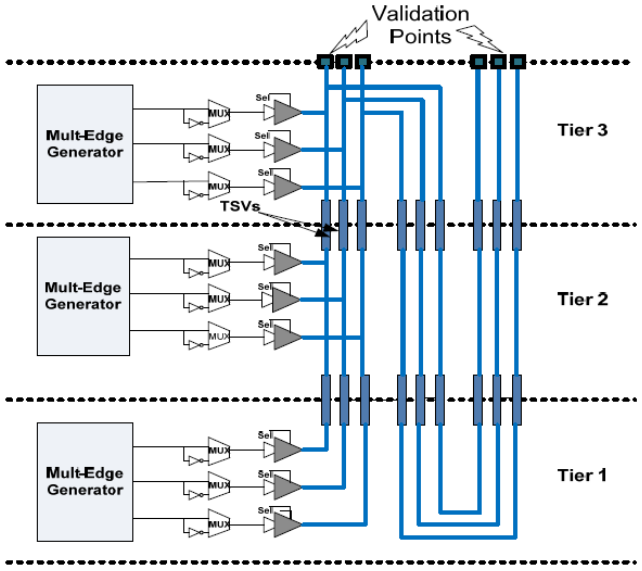


Figure 12. 3D Interconnect crosstalk induced-delay test

Figure 13 shows different crosstalk modes for studying crosstalk coupling on the victim edge. For example; -2X presents the case when both aggressors transition in the same direction as victim. -1X: when one aggressor transitions in same direction and other is quiet. 0X: when both aggressors are quiet or transition in opposite directions or the victim is quiet. +1X: when one aggressor transitions in opposite direction and other is quiet. +2X: when both aggressors transition in opposite direction as victim.

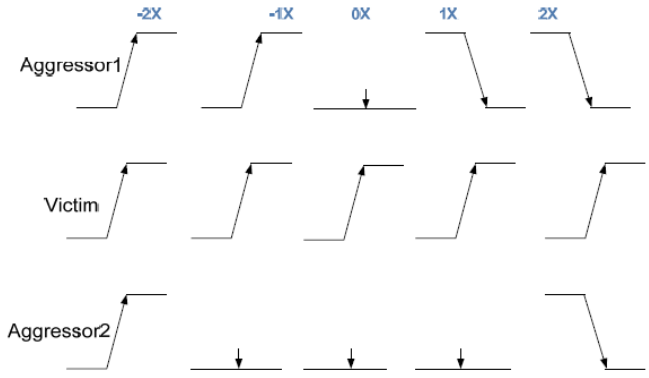


Figure 13. Different Crosstalk Patterns

Figure 14 shows simulated results for the induced-delay crosstalk with different patterns using ADS and 0.25 µm CMOS standard process. The middle graph is the victim line signal with no activity on both aggressor lines. The furthest right and left graphs are -2X and 2X cases respectively, which cause the worst case induced-delay effect. This induced-delay can be mitigated by phase shifting the aggressor signals using the multi-edge delay generator. Figure 15 shows an example of the crosstalk induced-delay cancellation effect after phase shifting the aggressor line 0.8ns for the case of +1X. As graphed, the induced delay due to crosstalk is almost cancelled and the signal aligns again with the 0X case (i.e. without crosstalk).

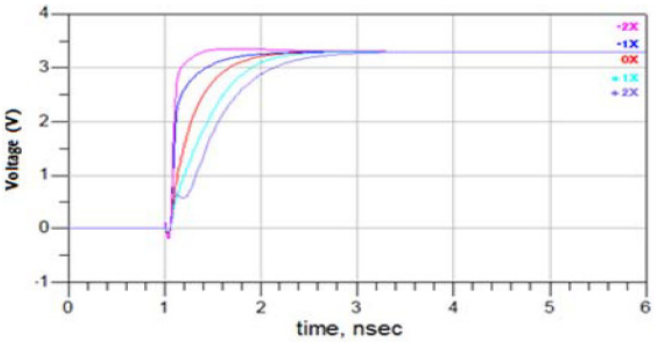


Figure 14. 3D interconnect induced-delay crosstalk for different crosstalk patterns

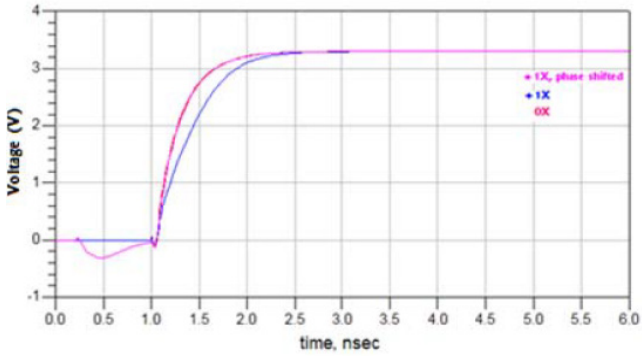


Figure 15. 3D interconnect induced-delay crosstalk cancellation

4.2. Ring Oscillators 3D Crosstalk Test

In this test, a consecutive triggered parallel Ring Oscillators (ROs) structure running same frequency is used to characterize the crosstalk effect between TSVs. Figure 16 shows four triggered oscillators; two oscillators are crosstalk-coupled and the other two are crosstalk-free. The proposed ROs parallel structure creates a delta phase shift difference between each consecutive triggered oscillators, which is equal to the time difference between the delay buffer chain and the oscillation time period of the triggered oscillators. 3D crosstalk detection can be achieved by observing the frequency of crosstalk-coupled oscillators, which is different from the frequency of the crosstalk-free oscillators.

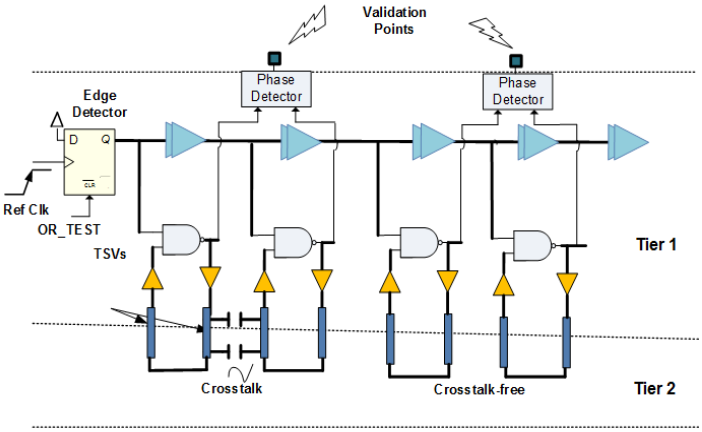


Figure 16: 3D Crosstalk Detection Scheme

Figure 17 shows the output frequency for both crosstalk-coupled (725 MHz) and crosstalk-free oscillators (667MHz).

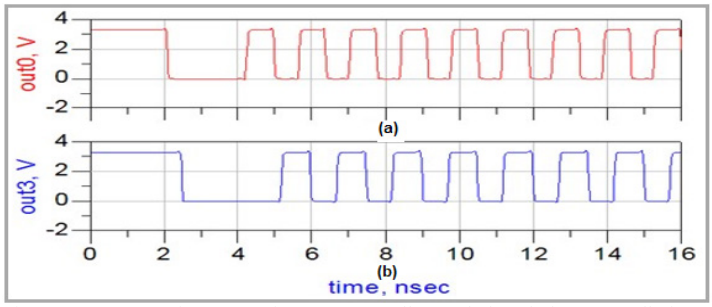


Figure 17: Output frequency for (a) Crosstalk-coupled and (b) Crosstalk-free oscillators

Interestingly, the two crosstalk-coupled triggered oscillators have less oscillation time delay (i.e. faster frequency) than the crosstalk-free oscillators, and the phase difference between the two crosstalk-coupled oscillators diminishes due to the coupling effect as shown in Figure 18. Phase detection at the output of the crosstalk-coupled triggered oscillators can be used as an indicator for strong coupling between TSVs. On the other hand, the edges of crosstalk-free oscillators are still separated by a deterministic phase shift dictated by the time difference between the delay buffer chain and the ring oscillation time period.

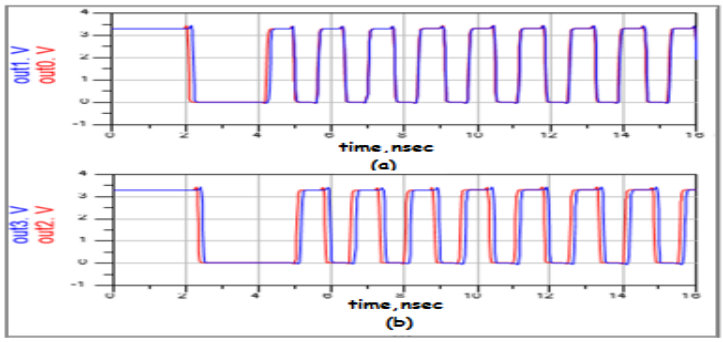


Figure 18: Simulated Results for 3D Interconnect Crosstalk (a) Crosstalk-coupled (b) Crosstalk free

5. Conclusion

In this paper, we presented the signal integrity effects of crosstalk in 3D stacked ICs. A detailed study of TSVs electrical modeling and characterization using HFSS and ADS simulators for frequency and time domains analysis respectively was presented. Simulation results were conducted based on 0.15µm 3DFDSOI process technology from MIT Lincoln lab which present the influence of 3D Vias distances, locations and their structural configurations on crosstalk. The study shows that increasing and decreasing distance of 3D Via signals and grounds respectively can mitigate significantly the effect of 3D crosstalk. In addition, adding more reference Vias and creating a well-defined return current paths have the most impact on mitigating crosstalk. Furthermore, it was shown that the discontinuity between 3D Via and transmission line has negligible impact on near-end crosstalk (NEXT), however; far-end crosstalk (FEXT) might increase but with less significant impact.

Furthermore, a 3D testing circuit application based on a multi-edge signal generator placed at different 3D stacked tiers was studied to evaluate the effect of crosstalk induced-delay and glitches. Additionally, a cross-coupled parallel ROs structure was

also presented to evaluate the crosstalk coupling strength effect compared to ROs structure with crosstalk-free TSVs.

References

[1] A. W. Topol, D. C. L. Tulipe, J. L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini and M. Ieong, "Three-dimensional integrated circuits," IBM Journal of Res. and Dev, vol. 50(4-5), pp. 491–506, 2006.

[2] J.-Q. Lu, "3-D hyperintegration and packaging technologies for micronano systems," Proc. IEEE, vol. 97, no. 1, pp. 18–30, Jan. 2009.

[3] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon, "Demystifying 3D ICs: the pros and cons of going vertical," *IEEE Design & Test of Computers*, vol. 22(6), pp.498–510, 2005

[4] J. Q. Lu, "3-D Hyperintegration and Packaging Technologies for Micro-Nano Systems," in Proc. of IEEE, vol. 97(1), 2009, pp. 18-30

[5] K. Tu, "Reliability challenges in 3D IC packaging technology," Microelectronics Reliability, vol. 51, pp. 517–523, 2011.

[6] T. Song, C. Liu, D. H. Kim, S. K. Lim, J. Cho, J. Kim, J. S. Pak, S. Ahn, J. Kim, and K. Yoon, "Analysis of TSV-to-TSV coupling with high-impedance termination in 3D ICs," in International Symposium on Quality Electronic Design, 2011.

[7] J. Tanskanen, J. Toikka, and E. O. Ristolainen, "Crosstalk of wiring in very small 3D module: IMAPS," 36th International Symposium of Microelectronics, pp. 251-255, 2003.

[8] T. Kang; Z. Yan; W. Zhang; J. Wang, "Research on crosstalk issue of through silicon via for 3D integration", *28th IEEE International Systemon-Chip Conference (SOCC)*, 2015.

[9] T. M. Mak, "Test challenges for 3D circuits," 12th IEEE International OnLine Testing Symposium (IOLTS), pp. 79–79, 2006.

[10] Shadi MS. Harb, William Eisenstadt, "A Study of Characterizing Crosstalk Effects in 3-D Vias," Accepted at the IEEE Latin American Symposium on Circuits and Systems Conference (LASCAS 2017), Bariloche,, Argentina, Feb 20, 2017

[11] MITLL Low-Power FDSOI CMOS Process Design Guide, MIT Lincoln Laboratory, 2008.

[12] Pedram S N M, "Crosstalk-affected delay analysis in nanometer technologies", International Journal of Electronics, Sept 2008.

[13] Gope D and Walker D M H, IEEE 30th International Conference on. IEEE," Maximizing crosstalk-induced slowdown during path delay test", p.159-166 , 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 115-124 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Constructing Learning-by-Doing Pedagogical Model for Delivering 21st Century Engineering Education

Ghassan Frache[1], Hector Nistazakis[2], George Tombras[*,2]

[1]*Abu Dhabi Vocational Education & Training Institute, United Arab Emirates*

[2] *Division of Electronics, Computers, Telecommunications and Control, Dept of Physics, National and Kapodistrian University of Athens, 15784 Athens, Greece*

A R T I C L E I N F O

Article history:
Received: 05 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
Higher education
Learning-by-Doing (LBD)
Pedagogical
21st Century skills
Pragmatic

A B S T R A C T

The United Arab Emirates (UAE) is dedicated to establishing the best teaching and learning environment for students and staff across all higher education institutions. To heed this call, the engineering division adopted the Learning-by-Doing (LBD) pedagogical philosophy for 21st Century education at the heart of its strategic directions. This study intends to explore how LBD is understood and practiced in UAE colleges and how 21st Century skills can be explicitly incorporated into its engineering curriculum by using constructive alignment as a pattern for instructional design. This work intends to investigate the general question: "What constructively aligned Learning-by-Doing pedagogical model, with the incorporation of 21st Century skills, needs to be developed to effectively teach and prepare engineering students at the Higher College of Technology, UAE for successful long-term employment in the global working economy?". Using mixed method research design and input from its major stakeholders, student survey questionnaires, engineering instructors and dean structured interviews, overt class observation and syllabi examination have all been utilized for data triangulation. In conclusion, the study developed a collaborative LBD model tailor-fitted to the institution's engineering program and to UAE's culture.

1. Introduction

This study of students' and instructors' perception of the implementation of Learning-by-Doing and 21st Century skills was conducted in Higher Colleges of Technology, HCT, a tertiary education institution in the United Arab Emirates (UAE). This research is an extension of work presented at the 2017 Global Engineering Education Conference [1]. Additionally, the work is part of thesis work for doctoral studies. The 2003 Emiratisation program, developed for the purpose of training Emiratis to become competitive, has been strictly enforced by His Highness Shaikh Mohammad Bin Rashid Al Maktoum, the Vice President and Prime Minister of the UAE (2012) [2]. The Emiratisation program is aligned with the UAE Vision 2021[3], which aims to develop and train Emiratis for taking up jobs in the country. Among other things, Vision 2021 foresees higher education as an environment where students will "enrich their minds with the skills that their nation needs to fuel its knowledge economy" [3, p. 16]. Thus, it can also be observed that part of Vision 2021's directive is for the UAE universities to "listen closely to the needs of Emiratis and of their future employers" and to "balance their teaching to the demands of the workplace" [3, p. 16]. Teaching methods and approaches need to be aligned with this requirement of matching the students' learning with practical needs of jobs in future, and Learning-By-Doing is one method that is gaining a reputation for doing just that [4]. This study is therefore conducted in a tertiary educational institution in the United Arab Emirates (UAE), with the aim of evaluating the pedagogical approach in the context of the framework of 'Learning-By-Doing'.

The concept of a more authentic, relevant learning has been a focus for educators since the time of John Dewey in the early part of the last century. Dewey's concept of "learning by doing" was based on his understanding that people learn best when they are actively involved in tasks that have meaning and importance [4]. In this last century, education has also been shaped by our growing understanding of how people learn. Johnson, Johnson and Holubec [5] argued that the work of Vygotsky in the 1920s, Jean

*Corresponding Author : G.S.Tombras, National and Kapodistrian University of Athens, 15784 Athens, Greece, gtombras@phys.uoa.gr

Piaget's cognitive development stages, Bloom and his now famous taxonomy, Howard Gardner's multiple intelligences and other related research studies have deepened our understanding of human learning, which, as [4] advocated should involve all aspects concerning the mind, hands and heart. Additionally, the principle of Learning-By-Doing is manifested in many of today's learning theories. Experiential learning, under which falls-active learning (whose subsets include cooperative learning and collaborative learning) and service learning – all exemplify the principle of Learning-By-Doing [6]. According to Voogt and Roblin [7, p. 29], in order for students to be successful in the 21st century, not only are certain skills necessary but these specific skills also need to be taught through (and are best supported by) "specific pedagogic techniques, such as problem-based learning, cooperative learning, experiential learning, and formative assessment", which again is encompassed in the approach of Learning-By-Doing [6].

The study aims to answer the overarching research question: What constructively aligned Learning-by-Doing pedagogical model, with the incorporation of 21st Century skills, needs to be developed to effectively teach and prepare engineering students at the Higher College of Technology, UAE for successful long-term employment in the global working economy? The main objectives of the research are as follows:

- To analyze the current understanding of LBD from the viewpoint of the engineering college's dean and instructors.
- To identify what LBD practices are being successfully implemented in the engineering department, from the perspectives of instructors and students in HCT.
- To identify what 21st Century skills are intentionally or incidentally taught and assessed in the practice of LBD, from the perspectives of instructors and students in HCT.
- To identify what pedagogical model might best meet the needs of implementing LBD and 21st Century skills in the engineering faculty at HCT.

The following sub-questions were formulated to guide the two phases of the study and provide conclusions:

RQ1. What are the current understandings of LBD from the viewpoint of the engineering dean and the instructors?

RQ2. From the perspectives of engineering instructors and students, what LBD practices are successfully implemented in the engineering department?

RQ3. From the perspectives of instructors and students, what 21st Century skills are taught and assessed in the practice of LBD?

RQ4. What pedagogical model might best meet the needs of implementing LBD and 21st Century skills in the engineering division at HCT?

2. Research Locale and Participants

The UAE is identified by PISA (Programme for International Student Assessment) as one of the most rapidly improving education systems in the world [9]. The UAE government recognizes that in 21st century economies, knowledge and skills have become so essential that a high value is placed on building a world-class education system that nurtures the minds and hearts of the UAE citizens. The government of UAE plays a dominant role in education, for schools and universities. It typically controls curricula even at private schools [9]. Most instructors are government employees, and most education is publicly financed up to the degree level. There are several types of post-secondary institutions in the UAE. A university offers degrees in several disciplines and usually offers graduate studies, while university colleges focus on one major discipline and are mostly for bachelor degrees. Technical institutes offer two- or three-year diplomas and are common throughout most of the region, accounting for one-third of all post-secondary students.

The UAE has more than 30 foreign academic institutions. The establishment of those institutes is a manifestation of the globalization of higher education and an indication of some of the UAE's attempts to become a knowledge-based society. Public universities in the UAE have separate colleges for males and females, following Islamic traditions. Such is the case with HCT, the locale of the present study.

HCT is one of the public's higher education institutions in the UAE, established in 2006. All the engineering programs are internationally recognized by accrediting bodies specifically the American Accreditation Board for Engineering and Technology (ABET). Because the proposed model is specific to the engineering department of HCT, the participants are its stakeholders: the engineering dean, the instructors teaching major courses in engineering and the engineering students. Only the instructors who have signed an informed consent to participate in the research were interviewed and observed. Table I shows the number of participants who were invited and participated in the study.

Table I: Participants of The Study

Colleges	Year 1	Year 2	Year 3	Year 4	Instructors	Dean
MZC Colleges (women)	18	15	0	0	2	
RUC colleges (women)	31	16			3	
RUC colleges (men)	30	41	32	18	3	
Total	79	72	32	18	8	1

3. Significance of the Study

The researcher believe that while LBD may be translated into various strategies that call for student-centered activities, an explicit definition based on shared understanding, contextualized to how LBD should be applied in the HCT colleges, would make its practice more focused. This has the potential to greatly contribute to achieving the college's mission of providing educational experiences that will infuse its graduates with "the knowledge, skills, and attributes to effectively contribute to the nation-building process and to help them develop a sense of personal and social responsibility". (http://www.hct.ac.ae/about/learning-model).

Additionally, the researcher sees the viewpoints of HCT major stakeholders as vital in developing explicit foundational commitments from the college community in lieu of one solely determined by higher management and handed down to teachers

for mandated implementation. Hence, the researcher takes into account the teachers, the students, and the management's viewpoints. It is anticipated these findings will contribute to the development of an explicit model that will help the teachers focus methodologies, assessments, and other pedagogical activities not only to the principal elements that comprise LBD but also for the promotion of learned 21st Century skills developed amongst students.

4. Implications for the Present Research

Although the literature does not provide any explicit definition of LBD, it links it to a number of important elements that ultimately promote its philosophy. This gives a more definitive idea of what it involves and a basis for investigating how the HCT instructor individually understands and practices the philosophy of LBD under the guidance of the college. The following is an analysis of what LBD involves:

- Learning to do (skills) not just to know, [10] - [11];
- Learning that is experiential [11], [12], active [10], [1],[13], [14], collaborative and cooperative [15];
- Learning that occurs in the context of a goal that is relevant and interesting to the student [2], [16];
- Learning that is planned (not discovered) [16], [2];
- Learning that involves not only quality academic reflection[17], [18] but civic (global) and personal experiences as well [18];
- Learning that considers the students' cultural context [19], [20] and respects every student's experience and builds upon these [11], [16];
- Learning that involves practical experiences in the context of relevant tasks closely related to how students will use it outside the learning environment [1], [5];
- Learning that involves strategies such as presentation, reports, team building, online contact time with students, critical thinking, studio teaching, team projects, and open-ended problem solving [20], [21] - [22];

The literature reviewed leaves no doubt as to the importance of 21st Century skills in today's education, particularly to engineering education. The review has also shown that a number of 21st Century skills take their roots from Dewey's (1899) work, among others.

Also, as mentioned in the research reviewed, over and beyond foundation knowledge taught through content engineering courses, the following are deemed indispensable 21st Century skills applied to an engineering education. The researcher adopts Mishra and Kereluik's re-categorization of 21st Century skills in presenting these.

1. Meta knowledge
   - critical and self-critical abilities/problem solving, [23] - [25];
   - communication and collaboration, [23] - [26];
   - teamwork [23] - [27];
   - creativity/innovation [23];
2. Humanistic knowledge
   - life and job skills [23], [25], [26]
   - ethics and cultural knowledge [26];

Interestingly, the LBD elements as listed above align with a number of the 21st Century skills reflected in the HCT learning model. To fully explore the alignment between LBD elements and 21st Century skills necessary in engineering education as outlined in the learning outcomes of HCT, a brief analysis by the researcher is presented in Table 2.

Table 2: Alignment Between LBD Elements, 21st and HCT Learning Model

LBD elements	21st Century skills	HCT learning outcomes
• Learning to do (skills) not just to know (factual knowledge); experiential, active, cooperative, collaborative learning	creativity/innovation life and job skills/communication and collaboration teamwork	Critical and creative thinking Vocational competencies Communication literacy Teamwork and leadership
• Learning that occurs in the context of a goal that is relevant, meaningful and interesting to the student	life and job skills communication critical and self-critical abilities/problem-solving	Vocational competencies Communication literacy Critical and creative thinking Self-management and independent Learning
• Learning that is planned	all skills should be planned learning outcomes	HCT learning model mandate
• Learning that involves quality reflection	critical and self-critical abilities ethics and cultural knowledge	Self-management and independent learning Global awareness and citizenship
• Learning that considers culture	ethics and cultural knowledge	Global awareness and citizenship
• Learning that involves practical experiences in the context of relevant tasks closely related to how students will use it outside the learning environment	creativity/innovation life and job skills communication and collaboration teamwork	Critical and creative thinking Vocational competencies Communication literacy Teamwork and leadership Mathematical literacy
• Learning that involves strategies such as presentation, reports, team building, critical thinking, steam projects, and problem-solving	communication and collaboration teamwork critical and self-critical abilities	Critical and creative thinking Teamwork and leadership Communication literacy Mathematical literacy

Biggs [28] instructional design, constructive alignment, aside from fully supporting LBD as it espouses experiential learning, sees the usefulness of an LBD structured system of learning in engineering education [29], [30]. The researcher finds this significant in contributing directions on how to go about model construction, giving particular focus to Biggs' alignment of learning outcomes, teaching and learning activities, and assessment.

In developing the model which this study proses to do, the researcher should keep in mind that although LBD is mandated as the teaching principle of Higher Colleges of Technology and is promoted in Biggs' instructional design [28], conventional methods of teaching still have a place in the classroom [23], [31] - [33]. Hence, the researcher notes that some lessons can be taught through interactive, high-level psychological lectures and activities rather than behavioral, experiential ones.

Additionally, following Houghton [34] and Fung [29] propositions, criterion-based assessments will be looked into while keeping in mind Voogt and Roblin's recommendation that while summative and formative assessments are useful in assessing 21st Century skills, new forms of assessment should build on previous assessment practices and should be considered as a starting point.

5. Research Methodology

As mentioned, HCT has mandated the use of LBD and 21st Century skills learning, having adopted these at the heart of its strategic directions. At this point, it is important to delve into how LBD is implemented and how 21st Century skills are infused especially from the collective perspectives of the different stakeholders – students, instructors, and the dean.

Keeping in view the pragmatic stance that reflects the belief that human experiences are multifaceted, and that agreement concerning what is and what should be in any given situation is best negotiated among all concerned parties, the mixed method approach was selected as the best-suited method for this study. The selection of the mixed method approach was in keeping with the stance of Venkatesh, Brown, and Bala [35] who state that "such work can help develop rich insights into various phenomena of interest that cannot be fully understood using only a quantitative or qualitative method" (p. 21). Creswell and Plano Clark [36] claim that using both qualitative and quantitative methods enables the researcher to reach a more comprehensive response to the research questions, as opposed to the possible limitation of using only one method. Using the mixed methods design in this research has enabled reliable and valid data to be elicited, as it helped the researcher in obtaining data from different sources using different approaches – in-depth interviews, observations, quantitative surveys, and document analysis. Migiro and Magangi [37] state that there are three important advantages of this method:

(1) A researcher can use strengths of an additional method to overcome the weaknesses in another method by using both in a research study;

(2) Mixed method research can provide stronger evidence for a conclusion through convergence and corroboration of findings;

(3) Mixed method research can add insight and understanding that might be missed when only a single method is used (p. 3763).

Hussein [38] states that using multiple ways of data collection in the same study increases the credibility of the study (p. 1). He asserts that using mixed methods achieves more accurate and reliable findings with richer information.

The use of mixed methods is therefore grounded in the researcher's belief that findings from each set of participants (students, teachers, and the dean of engineering) will give not only a richer depiction of the findings but also a unique description of what LBD practices and 21st Century skills are currently implemented at HCT.

Quantitative Tools

Quantitative methods are often used in social science to acquire knowledge by manipulating data through sophisticated quantitative approaches, such as multivariate statistical analysis [22]. The engineering students were requested to repond using Likert-scale questions by employing an online survey framework that was used by the researcher. Keeping in mind participant fatigue, the questionnaires were divided into two sets: LBD elements and 21st Century skills. The sets were administered separately, a week after the other.

Qualitative Tools

The qualitative data collection was selected in order to offer the researcher an in-depth perspective regarding the personal experiences. Reference [23] provided a succinct definition that "qualitative research seeks to discover new knowledge by retaining complexities as they exist in natural settings." Reference [24] provides a description of a variety of data collection tools used in qualitative studies: interviews; data analysis; direct observations; and reporting. The choice of tool is influenced by the skill of the researcher, data collection strategy, the type of variable, the accuracy required, and the collection point [21].

Data Analysis

The following data analysis approach was shaped by the type and amount of data collected and made extensive use of description in addition to a number of statistical analyses of the quantitative data. The primary data sources, survey questionnaires, semi-structured interviews, content analysis and over classroom observations, were analysed and reported in terms of patterns in participants' responses in relation to the main and enabling research questions that guided the study.

1) Quantitative Data Analysis

To test whether the actual values are significantly different from the expected values, the chi-square test was applied to the LBD and 21st Century questionnaire. A null hypothesis signifies that there is no statistically significant difference and an alternate hypothesis states the opposite. Based on the results of the above

test, the researcher can either reject or fail to reject the null hypothesis. The techniques and methods used in this study encompass the following: The descriptive analysis by using the mean and standard deviations was conducted for the items determining the 21st Century skills.

2) Qualitative Data Analyses

To analyse the qualitative data, the audio recordings of the interviews and transcripts of the recordings were entered into nVivo software for coding and further analysis.

6. Data Analysis and Results

This section discusses the findings stemming from the quantitative and qualitative data collected. The quantitative methodology tools used are questionnaires given to one particular type of stakeholders, specifically the engineering students. A descriptive analysis was carried by using the frequencies and proportions. The Likert-scaled data of Learning-by-Doing items were tested for statistical significance by using a Chi-squared test. A descriptive analysis using the frequencies and proportions was conducted for the items determining the 21st Century skills. The qualitative methodology tools used are semi-structured, one-on-one interviews among the two sets of target participants: the engineering instructors and the dean.

A. Descriptive statistic Learning-by-Doing survey findings

The first questionnaire was administered to the engineering students seeking to understand how LBD is manifested inside the classroom. Students were given a set of 16 statements and then asked to rate the frequency of these statements. The statements were geared to correspond to LBD principles as opposed to LBD techniques. The survey questionnaire was administered to 184 engineering students. Students' ratings of each of the statements are listed in Table 3. Frequencies, Means, and Standard Deviation data analysis are shown. In addition, a graphical representation for the rating of LBD statement is shown in figure 1.

The first statement provides emphasis on activities done in the classroom which emphasizes collaboration as its main ingredient. Only 26% of the participants stated that this is always the case with 32% stating this is usually done. The second statement aims to find out whether there is a meaningful interaction between students and instructors during discussions. 32% stating this is always done. The third statement emphasizes reflection on activities conducted. This is one of the hallmarks of Learning-by-Doing. More than 50% of the students said it was done at least 4 times in six situations. The fourth and fifth statements are related to each other as they tackle the issue of learning "life skills," another characteristic of LBD. We can safely say more than 50% of students are saying these are being done 4 out of 6 situations. The 6th statement's aim was to ferret out whether "Active Learning" is happening in the classroom. This statement garnered the lowest number 5 rating (12%). The 7th statement refers to the use of simulation in the classroom. The results on this question are encouraging considering it has a high rate of "Usually" (38%). The demonstration of a concept is the emphasis of the 8th statement. Looking at the results, it can be seen that it has the highest points garnered in the data spectrum of "Always" (30%). The 9th statement belongs to the purview of "Active Learning," conducting drills and practice. The results show that this statement has the highest frequency number for the rate of "Usually," garnering a whopping 44%. The 10th statement is part of the reflection phase of LBD where students are supposed to retrospect on what they have learned. The combination of the "Usually" and "Always" frequency score however still ensures that more than 50% of the participants believe that this happens. The 11th statement touches the post-evaluative aspect. LBD encourages different ways of evaluation to encourage the participation of the learner or learners in this phase. We can still conclude that more than 50% of students believe that this happens 4 times out of 6 situations. The 12th statement is part of the real-life learning of LBD were activities that encourage more "learner" immersion is an integral part. Suggested by the statement itself are workshops and field trips. This statement got the highest "Never" and "Almost Never" frequency score. The 13th statement is still within the context of real-life learning. The 14th statement is within the purview of "Active Learning," with at least 50% of the participants believing that this occurs in 4 out of 6 situations. The 15th statement is within the coverage of post-evaluation and reflection. Students are asked to record their impressions on a phase by phase basis. Almost 60% of the participants believe that this happens in at least 4 out of 6 situations. The 16th statement focuses on hand-ons, kinesthetic activities. This statement got the highest "Occasionally" rate, having 40%.

Table 3: Summary of Results for Each LBD Statement

	LBD statements	Frequency %					Sample Size	Mean	SD.P
		Always	Usually	Occasionally	Almost never	Never			
1	There are classroom activities that require students to **collaborate** and learn with and from each other. Examples are **group projects** that emphasize **teamwork**.	26	32	26	12	4	184	20	10.2
2	Discussions in the classroom are **interactive** meaning students as well as the teacher contribute to the topic being discussed.	32	32	25	9	3	184	20	12.0
3	Questions and answers that focus o**n post-evaluation** of learning activities are conducted in the sessions.	25	33	27	10	5	184	20	10.7
4	Exam questions are focused on **scenarios** that require students to apply what they have learned and are not merely limited to ones that call for memorization, definitions, etc.	21	34	30	10	4	184	20	11.4
5	The teacher uses **real life** case studies as a means for teaching the content of the course.	20	32	33	11	5	184	20	11.1
6	Students are presented with **problem-based** questions where students either in group or individually will work out the solutions.	12	37	35	12	4	184	20	13.4
7	Teacher uses **simulation** either digital or manual as a means of teaching a concept.	22	38	30	7	3	184	20	13.3
8	Teacher **demonstrates** a required subject skill first then asks the students to follow suit.	30	35	26	7	2	184	20	13.1
9	Students **do drills and practice** as a means of learning and **mastering a skill** or a concept.	29	44	19	7	2	184	20	15.1
10	Students are encouraged to **reflect** on what they have learned and express this reflection either orally or in written format.	21	34	33	9	3	184	20	12.5
11	In assessing student's work, the teacher uses other means in addition to his/her own assessment. This other means can be **self-assessment or peer review.**	15	36	30	11	7	184	20	11.2
12	Teachers conduct activities that allow students to fully experience the topic. Examples of these type of activities are **field trips and workshops.**	16	29	33	14	8	184	20	9.4
13	The college provides programs that bring students to the **workplace** as part of the student's preparation for professional working life after graduation.	15	34	32	11	8	184	20	10.8
14	Classroom activities that ask the students to **model experiences** or concepts are conducted. Examples of these types of activities are **role-playing,** reenactment or walkthrough (From process to output)	17	33	35	9	5	184	20	12.2
15	Teachers encourage students to **record their impressions** on how they made the project on a **phase-by-phase** basis. This requirement is in addition to the required output of the project.	29	31	26	11	3	184	20	11.1
16	Classroom activities are formulated in such a way that students can be more **active and motivated** in their work. Examples of this type of activity are educational games and other hands-on means.	23	28	36	9	4	184	20	11.9

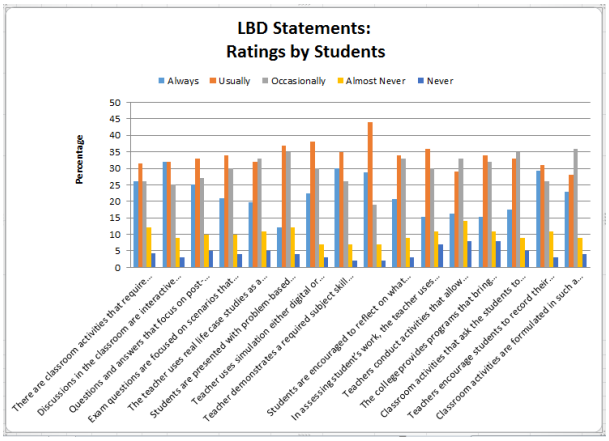


Figure 1. LBD Statements

B. Chi-Square Test on Learning-by-Doing statements

In order to determine whether there is a significant difference between the opinions of the participants regarding the LBD statements, a chi-square test for equal proportions was applied. The null and alternate hypotheses are as follows:

Null hypothesis: There is no significant difference between the opinions of the participants regarding each statement.

Alternate hypothesis: There is a significant difference between the opinions of the participants regarding each statement.

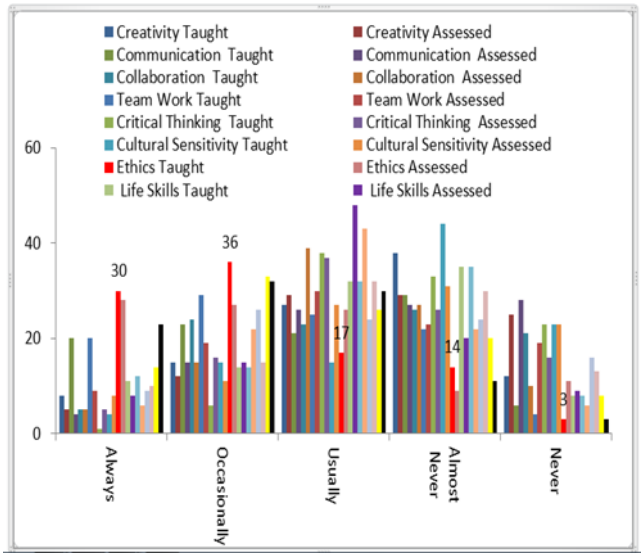
Since the p-value is less than 0.05, we can conclude that there is little significant difference between the participants' opinions.

C. 21stCentury skills survey findings

To answer the 3rd question of the study, the student's point of view with regard to whether 21st Century skills are either intentionally or inadvertently taught or assessed in the classroom was addressed. The manifestation of the skills was established in the questionnaire to crystalize the concept in the mind of the students. Students rated each category based on a 1 to 5 frequency scale. Findings of the survey results are shown in Table 4.

Table 4: Summary of Findings for 21st Century Skills.

	21st Century Skills		Frequency %					Sample Size	Mean	SD.P
			Always	Occasionally	Usually	Almost Never	Never			
1	Creativity	Taught	8	15	27	38	12	132	20	11.0
		Assessed	5	12	29	29	25	132	20	9.8
2	Communication	Taught	20	23	21	29	6	132	19.8	7.6
		Assessed	4	15	26	27	28	132	20	9.3
3	Collaboration	Taught	5	24	23	26	21	132	19.8	7.6
		Assessed	5	15	39	27	10	132	19.2	12.3
4	Team Work	Taught	20	29	25	22	4	132	20	8.6
		Assessed	9	19	30	23	19	132	20	6.8
5	Critical Thinking	Taught	1	6	38	33	23	132	20.2	14.6
		Assessed	5	16	37	26	16	130	20	10.8
6	Cultural Sensitivity	Taught	4	15	15	44	23	128	20.2	13.3
		Assessed	8	11	27	31	23	132	20	9.0
7	Ethics	Taught	30	36	17	14	3	132	20	11.7
		Assessed	28	27	26	9	11	132	20.2	8.4
8	Life Skills	Taught	11	14	32	35	8	132	20	11.2
		Assessed	8	15	48	20	9	131	20	14.7
9	Problem Solving	Taught	12	14	32	35	8	130	20.2	11.1
		Assessed	6	22	43	22	6	129	19.8	13.6
10	Innovation	Taught	9	26	24	24	16	131	19.8	6.4
		Assessed	10	15	32	30	13	131	20	9.1
11	Computer Literacy	Taught	14	33	26	20	8	132	20.2	8.8
		Assessed	23	32	30	11	3	132	19.8	11.2



Creativity: It is observed that 50% of the students think creativity as a skill is being taught and assessed only once in 6 situations. It is interesting to note that about 54% of the participants graded 1 (never) or 2 (almost never) for they now think outside the box by exploring alternative ideas and solutions. It is interesting to note that creativity taught and assessed received 8% and 5% respectively. This shows very little emphasis is devoted to creativity in the classroom by the teachers.

Communication: Communication skills reached the 43% in terms of students believing that it happens 4 out of 6 times in the category of taught and assessed. It would seem that communication is another skill that is not well assessed with 55% of the students thinking that never/almost never done is happening here.

Collaboration: Looking at the taught and assessed categories, we can see that only 5% of the participants think this is done 6 times out of 6 situations and 44% think this is done 4 out of 6 situations. We can conclude that collaboration is not taught nor assessed in the classroom

Team work: Can also be considered as one of the top skills in terms of scoring as it has a more than 40% rating in the "Usually" and "Always" categories. Just like in the other skills it means 50% of the students believe it is happening 4 out of 6 times.

Critical thinking: This skill scored lowe in both categories. These results support the obtained data from the instructor interviews about the teaching and assessment of the critical thinking skill. It is observed that 56% of the students think critical thinking as a skill is being taught and assessed only 1 time in 6 situations.

Cultural sensitivity: This skill received the highest score of 44% for "Almost Never." This highlights that almost half of the participants think that cultural sensitivity is rarely being taught in class.

Ethics: Ethics is one of the top skills that the student consistently rated high. It has 30%, "Always" score based on the taught category, achieving the highest score among all the skills.

Life Skills: Life skills received an average score in both categories. In the taught category Life Skills received 14% in "Usually" and 11% in "Always."

Problem Solving: The problem-solving scores 12% in the taught and 14% that this happens 4 times out of 6 situations. This is surprising result given that most of the exercises in the engineering department require laboratory and math problems. It is observed that about 35% of the participants rated "Occasionally" for classroom activities that require students to examine different processes or paradigms and apply these to different types of problems.

Innovation: Similar to critical thinking and creativity, innovation received a low percentage in "Always." It has 9% in the taught category and 10% percent in assessed category.

Computer Literacy: In the category of "taught" alone, the combination of "Usually" and "Always" is 47% which means almost 50% of the participants agree that it happens at least 4 out of 6 times.

D. Learning-by-Doing as explained by participants

The study presents a discussion on the analysis of the results obtained from the primary research. The study aimed at evaluating the understanding of LBD at Higher Colleges of Technology (HCT). UAE has taken evident measures to enhance the quality of education, in order to prove its educational capabilities on the global platform. Hence, in order to assess if the government is successful in its endeavor, it was important to assess how HCT approaches and implements LBD and 21st Century skills. The research therefore aimed at understanding, from the perspectives of the teachers, students, and the dean of the Higher Colleges of Technology what constitutes of LBD and 21st Century skills, and which activities were in the process of implementation under LBD and 21st Century skills. This is part of primary research that was undertaken by conducting using semi-structured interviews (of teachers and the dean) and surveys (of students). In addition, eight classroom observations were also used to assess first-hand the activities employed in classroom and the approach of teachers while teaching. The classroom observations were then analyzed in conjunction with the interview and the survey findings. In addition, an exhaustive analysis of the HCT curriculum was also undertaken in order to review the LBD and 21st Century related content and activities present in it.

The participants' answers can be grouped into four different classifications; mainly, the practical aspect, real world impact, and definition of LBD. From the data gathered, it can be inferred that the participants have diverse views on Learning-By-Doing practices, and it is stated to have a different meaning depending upon the type of course(s) the faculty is teaching and prior professional experience of the participant educator. Table 5 reflects the answers of the participants when asked about their understanding of LBD.

In the concept of engineering, it was claimed by one of the participants that LBD is more about working on "analytical thinking, thinking and formulating a problem, and understanding the practical application of concepts." Another participant stated that, he looks at Learning-by-Doing as a "fairly permissive term and that it is primarily focused on experiential learning." The same view was shared by the engineering dean where he stated "my understanding on learning-by-doing is it's a fairly lenient term. So the first thing I would say is it hasn't got an exact definition."

Table 5. Summary of findings of LBD understanding

Findings	Occurrence
Different meanings in different context	1
Includes critical thinking	1
Laboratory work	1
Relevant learning experience	1
Embrace all other types of pedagogy	1
Hands on applicability	1
Old concept	1
Practical aspect of learning	4
Integral part of engineering	3
Supports theory	2
Relating to real world	4
Increase productivity of students	1
Involves careful design of activities that allows good foundation for knowledge.	1
Provides an opportunity to explore and develop new ideas.	1
Problem solving	1
Practical aspect	4
Real world	4
Impact of LBD	5
Definition of LBD	2

Almost all educator participants suggested during the interviews the need for a clearer and concise definition of LBD to be produced by the institution. Specifically, as one participant stated - "we need a more expansive definition of Learning-By-Doing that captures what distinguishes as well as what unites all members of engineering divisions in a shared educational enterprise." Table 6 captures the common understandings of LBD highlighted during the interviews.

Table 6. LBD defined

Findings	Frequency	Percentage
Experiential learning	2	25%
Life-long learning	1	12%
Problem solving	1	12%
Applying theoretical knowledge	3	37.5%
Apply learning to real life	3	37.5%
Project-based	1	12%

Almost all participants highlighted the importance of the union of theory and practice with an emphasis on the later. The view of marrying both theory and practice is not something new as it has been part of the LBD research literature tradition. For instance, Dewey [25] suggested the move from formal, abstract education to one that is more experienced-based. At the core of Dewey's LBD philosophy is action. Rather than merely thinking about abstract concepts, LBD involves a direct encounter with the phenomenon being studied. It utilizes actual experience with the phenomenon to validate a theory or concept [25]. It should be emphasized and pointed out that most of the participants believed that LBD is not a "stand-alone" philosophy or practice but instead intrinsically tied in with making the theory more relevant and understandable.

The experiential learning definition received the second highest percentage. This is well in agreement with widely

published literature on LBD and in its roots of formulation as a concept and action. Experiential learning and its subsets, cooperative learning and collaborative learning, exemplify the philosophy of LBD [26]. However, as verified in the interview and follow up questions, most of the participants understood experiential learning as a type of "hands-on" learning that does not really emphasize the promotion of discussion, critical thinking, reflection, introspection, and retrospection. In this aspect, the experiential learning that most of the participants subscribe to has something to do with involving the student in the learning experience in order for the learner to understand the concept, thus this is more akin to active learning [13].

Participants were asked, "What LBD activities do you do or have you done to teach your students?" Diverse practices were encountered in classrooms to achieve LBD. Table 7 summarizes these findings.

Table 7: LBD Practices

LBD Practices	Frequency	%
Laboratory activities/ experiment	5	21.7
Case study	1	4.3
Project-based	3	13
Building	1	4.3
Industry visit	2	8.6
Design process	1	4.3
Demonstration	5	21.7
Peer Demonstration	1	4.3
Exploration	1	4.3
Modeling and non-routine problems	1	4.3
Problem-based	2	8.6

The highest observed practices are laboratory activities and demonstration with 21.7% each. With respect to hands-on courses, the instructors are observed to assess the students and train them in LBD through practical experiments and labs. Certain other answers inferred that experiential learning is not only experiental or lab work. "I think having a reflective analysis and self-criticism or guided criticism of that experiential learning exercise is something that is often missed," the dean of engineering noted. It was also inferred that one of the biggest challenges is the understanding that pedagogy is the philosophy.

E. Participants view on 21st Centure practices

Participants were asked to answer several questions on how 21st Century skills were adopted in their classroom activities. The engineering dean said, "UAE engineering students need to possess them in order to survive in the current corporate world." He used the term "professional skills" for 21st century education to highlight those listed skills. Ongoing research findings reveal that to answer today's challenges, students must have the capacity to apply knowledge in practice by learning to adapt to new situations; critical and self-critical abilities; the ability to use teamwork and communication skills, with these being listed as the top three competencies needed by 21st century engineers [27]. The dean went on to highlight the opportunities and challenges of engineering education in the 21st century, contending that the new professional engineer not only needs to be knowledgeable in his own discipline but also needs a new set of professional skills. Table 8 summaries the 21st Century skills findings.

Table 8: 21st Century Skills Findings

Participant	21st Century skills understanding	21st Century skills used	Assessed
Dean	The participant has a clear understanding of 21st Century skills. He even defined some of the skills to include theirelements in the assessment of students.	50% of courses use the 21st Century skills.	Not really, but should be included with clear rubric during lesson and assessment.
Teacher 1	It seems that the interviewee is unsure of the definition of those skills. He struggles to distinguish between terms such as creativity and innovation.	Innovation, creativity, team work, problem-solving, communicati on, ethics.	It does not seem that the teacher is aware of how to assess his students on any 21st Century skills. None of these skills are clearly assessed.
Teacher 2	It seems the participant is unclear on the definition of 21st Century skills.	Team Work, collaboration , and cultural sensitivity.	Yes, informally. Not with a set of clear rubric.
Teacher 3	The participant was aware of the 21st Century skills and answered some questions about their definitions	Creativity, collaboration , teamwork, critical thinking, problem-solving and innovation.	Yes. In a rubric. Official rubric tool was supposed to be presented to the interviewees.
Teacher 4	"I don't understand what you mean by communication." She seems to understand all the other skills.	Problem-solving, critical thinking, teamwork, collaboration , and life-skills.	Not all are assessed. Problem-solving, critical thinking, teamwork, and collaboration are hard to assess. But others I do assess in LBD activities.
Teacher 5	Don't understand the cultural sensitivity. Some confusion about the definitions of some skills.	All skills are used in LBD activities.	All skills are assessed except ethics and critical thinking. No rubric is available.
Teacher 6	The participant is clear about the definition of the 21st Century skills.	Creativity, communicati on, collaboration , teamwork.	All are assessed. Some might have an indirect assessment. No rubric presented.
Teacher 7	The participant seems to understand 21st Century skills.	Teamwork, collaboration almost all of the 21st Century skills.	All indirectly assessed. No rubric is available.
Teacher 8	The participant was not aware of the distinction between innovation and creativity. Also, he did not understand the use of life skills in classrooms.	Collaboratio n, team work, ICT and life cultural sensitivity is used.	None of the skills are assessed directly.

It is obvious that not all participants understood what 21st Century skills are, let alone what the definition is or what the included activites are. This is despite the fact that a working definition of the skills was sent to them before they started teaching their courses at the beginning of the semester. While the participants understood the general meaning of the terms, they are at a loss when they are asked about them in the context of their inclusion in classroom activities and assessments. This is not surprising since most of the assessments of engineering subjects are technical in nature which means they usually use a quantitative approach. Skills such as collaboration, creativity, innovation, ethics, and cultural sensitivity would seem to require a qualitative evaluation framework in order to be assessed. 90% of the participants admitted that they do not include a majority of these skills as part of their assessment. When pressed with a follow-up question on the reason why, most of them simply said it was not required in the course outline. Some ventured to say it is difficult to assess these skills as they seem to be fraught with subjectivity which to them is a departure to the objective and clear type of assessment engineering students are used to having. However, not all 21st Century skills were unfamiliar to the participants. A clear majority of them stated that technological literacy, problem-solving, teamwork, ethics, and collaboration are mostly included in their assessment tools.

F. View of the participants towards incorporating the Learning-By-Doing and 21st Century skills better in the course outline

The participants also provided their views on how LBD and 21st Century skills were incorporated in the course outline. Mostly, it was stated that the course outline should include clear assessment guidelines for LBD activities and 21st Century skills. Two of the participants emphasized the core concepts and skills students needed to learn and cautioned to avoid unconnected topics which inhibit the development of critical thinking and other 21st Century skills. In addition, the participants also claimed that critical and creative thinking can be incorporated into the course outline, and on giving more importance to improving the communication abilities and promoting the spirit of teamwork and leadership. One of the participants specified four main skill areas that affect creativity and innovation in the current course outline: *fluency, flexibility, elaboration and originality*. Hence it was proposed that making these skills a part of the assessment strategy would better incorporate LBD and 21st Century skills in the course outline. The participants also claimed that the personal and social development of the students may be enhanced by developing their managerial and leadership skills, and by preparing them to implement complex skills such as planning, communicating, problem-solving, and decision-making.

Conclusion

This study reported on data gathered from instructors and the HCT dean's interviews and student questionnaire surveys. It has shown diverse views on the Learning-by-Doing understanding, practices, and definition. While most of the educator and student participants explained their understanding of LBD, the enumeration of the impact verifies the "incomplete" understanding of LBD. On the definition of LBD, participants viewed it as the second half of "theoretical" knowledge which is its application. This understanding is supported by LBD research literature. For instance, [7] suggested the move from formal education to one that is more experienced-based. As engineering is a skill-based field, it was no surprise for the participants to award high percentage for laboratory activities and experiments to be the most used LBD activities. The importance of 21st Century skills was highlighted by all participants during the educator participant interviews. But, most of them failed to use them in the context of their classroom activities. Creativity, critical thinking and innovation skills were highlighted as the most difficult to assess in engineering courses. The underlying theme of the findings from the students' point of view is that nearly 55% of the students believed the LBD is conducted and assessed in the classroom. The rest of the students, 45% seem to believe that LBD is not conducted in the classroom activities. The survey data used in this study was discussed along with other interviews and personal observation data to explore the results further in the construction of the LBD model.

References

[1] G. Frache; H. E. Nistazakis; G. S. Tombras. "Reengineering engineering education: Developing a constructively aligned learning-by-doing pedagogical model for 21st Century education", 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, Greece, 2017, pp. 1119 – 1124

Doi: 10.1109/EDUCON.2017.7942989.

[2] Zaman, S. (2012). 2013 dedicated to Emiratisation. *Gulf News*. Retrieved from http://gulfnews.com/news/gulf/uae/society/2013-dedicated-to-emiratisation-1.1111015.

[3] Vision 2021 United Arab Emirates. Available: http://www.vision2021.ae/en

[4] J. Dewey, *Democracy and education: An introduction to the philosophy of education*. New York: WLC Books, 2009.

[5] D. W. Johnson, R. T. Johnson, and E. Holubec, *Cooperation in the classroom* (8th ed.). Edina, MN: Interaction Book Company, 2008.

[6] T. T. Pham. "Issues to consider when implementing student-centered learning practices at Asian higher education institutions". *Journal Of Higher Education Policy And Management*, 33 (5), pp. 519—528, 2011.

[7] J. Voogt, & N. Roblin, *International Symposium on Education Reform | EDL*. [online]. [Viewed: 2013]: ttp://www.internationalsymposiumoneducationalreform.com

[8] Organization for Economic Cooperation and Development. (2014). Better Skills, Better Jobs, Better Lives: A Strategic Approach to Education and Skills Policies for the United Arab Emirates. Paris, France.

[9] W. Rugh, "Arab education: Tradition, growth, and reform, "*The MiddleEast Journal*, 396-414, 2002.

[10] L. H. Lewis, and C. J. Williams, "Experiential learning: Past and present, " *New Directions For Adult And Continuing Education*, vol. 62, pp. 5-16, 1994

[11] A. Y. Kolb and D. A. Kolb, "Learning styles and learning spaces: Enhancing experiential learning in higher education, " Academy *Of Management Learning & Education*, vol. 4, no. 2, pp. 193-212, 2005.

[12] A. Finkel, "*Innovative Approaches to Engineering Education: The Australian Experience," 2013,* www.caets.org.

[13] R. Pascual, and R. Uribe, "Experiential Learning Strategies in a Mechanical Engineering Senior Course, ", paper presented at *Sixth International Workshop on Active Learning in Engineering Education*. Mexico: Monterrey, 2006.

[14] M. Romi, "Learning by teaching in engineering: a step beyond learning by doing, "Technology*, Education And Development, A. Lazinika and C. Calafate, Eds. Croatia: Intech*, 2009, pp. 337-394.

[15] P. R. Donald and J. L. Faust, "Techniques for active learning", 2010.

[16] L. Kane, "Educators, learners and active learning methodologies", Int. *Journal Of Lifelong Education*, vol. 23 no. 3, 2004.

[17] S. L. Ash and P. H. Clayton, "The articulated learning: An approach to guided reflection and assessment, " *Innovative Higher Education*, vol. 29 no. 2, pp. 137-154, 2004.

[18] R. C. Clark and R. E. Mayer, "Learning by viewing versus learning by doing: Evidence-based guidelines for principled learning environments,

"Performance *Improvement*, vol. 47 no. 9, pp. 5-13, 2008.

[19] M. Ginsburg, *Active-learning pedagogies as a reform initiative: Synthesis of case studies*. Washington, DC: Academy for Educational Development, 2009.

[20] R. Pascual and R. Uribe, "Experiential Learning Strategies in a Mechanical Engineering Senior Course", paper presented at *Sixth International Workshop on Active Learning in Engineering Education*. Monterrey, Mexico, June 2006.

[21] C. M. Vest, "Educating engineers for 2020 and beyond, "*Report of a Convocation "Educating the engineer of* 2020. pp. 160—169, 2005.

[22] P. Morgado, "From Passive to Active Learners: Implementing the Pedagogy of " Learning by Doing" in a Large-sized Design Foundation Class, "Transformative *Dialogues: Teaching & Learning Journal*, v.4 n.2, 2010.

[23] C. M.Vest, C. M, "Educating engineers for 2020 and beyond," *Report of a Convocation "Educating the engineer of* 2020. pp. 160—169, 2005.

[24] L. Morell, "Engineering Education in the 21st Century: Roles, Opportunities, and Challenges," *International Journal of Technology and Engineering Education, 7*(2), 1-10, 2010.

[25] National Academy of Engineering (NAE) (2010) Grand Challenges for Engineering in the Eyes of 21st Century Students. Available from http://www.edstechnologies.com

[26] K. E.Wolfe Understanding the careers of the alumni of the MIT Mechanical Engineering department, 2004.

[27] R. Pascual and R. Uribe, "Experiential Learning Strategies in a Mechanical Engineering Senior Course", paper presented at *Sixth International Workshop on Active Learning in Engineering Education*. Monterrey, Mexico, June 7-9, 2006.

[28] J. B. Biggs, C. S. Tang, and J. B. Biggs *Teaching for quality learning at university*. Maidenhead: Mcgraw-Hill/Society For Research Into Higher Education & Open University Press, 2007

[29] S. Nightingale and A. Carew and J. Fung, "*Application of constructive alignment principles to engineering education: Have we really changed*," In Proceedings of the Australasian Association of Engineering Education Conference, December, pp. 9-13, 2007

[30] N.M. Meyers and D. D. Nulty, "How to use (five) curriculum design principles to align authentic learning environments, assessment, students' approach to thinking and learning outcomes," *Assessment and Evaluation in Higher Education* (34), pp 565-577, 2009

[31] R. Clark and R. E. Mayer, "Learning by Viewing versus Learning by Doing: Evidence-Based Guidelines for Principled Learning Environments," *Performance Improvement*, *47*(9), 5-13, 2008

[32] L. Kane, "Educators, learners and active Learning Methodologies," *International Journal of Lifelong Education, 23*(3), 275-286, 2007.

[33] D. Hung and S. Lee, "Is there an instructional framework for the 21st century?" *Creative Education,* 3(4), 461-470, 2012.

[34] W. Houghton, *Engineering Subject Centre Guide: Learning and Teaching Theory for Engineering Academics*. Loughborough: HEA Engineering Subject Centre, 2004.

[35] V. Venkatesh and S. A. Brown, and H. Bala, "Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems," *MIS Quarterly*, *37*(1), 21-54, 2013.

[36] J. W. Creswell, and V. L. Clark, *Designing and conducting mixed methods research*. Thousand Oaks, Calif.: SAGE Publications, 2011.

[37] S. O. Migiro and B. A. Magangi, "Mixed methods: A review of literature and the future of the new research paradigm." *African Journal of BusinessManagement. 5*(10), 3757-3764. doi: 10.5897/AJBM09.082, 2011.

[38] A. Hussein, "The use of triangulation in social sciences research: Can qualitative and quantitative methods be combined, "*Journal of Comparative Social Work*, *1*, 1-12, 2009.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 125-129 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A 3D Full Wave Inversion (FWI) Analysis for Handheld Ground Penetrating Radar (GPR)

Suki Dauda Sule*, Kevin Paulson

University of Hull, School of Engineering and Computer Science, HU6 7RX, United Kingdom

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
GPR
Handheld
Full wave inversion

ABSTRACT

This paper reports the results of an empirical, 3D full wave inversion (FWI) numerical analysis which provides an estimation of the performance of multi-static handheld ground penetrating radar (GPR,) compared to a bi-static system, for landmine detection using FWI imaging. The experiments are based on simulated data and provide a more realistic evaluation of the performance of multi-static handheld GPR for the landmine problem based on FWI, than previous studies based on a 2D analysis.. Furthermore, a novel method of estimating a parameter set that is closer to the global minimum for an iterative FWI optimization is introduced.

1. Introduction

This paper provides an extension of the work presented in [1], which is based on a sensitivity analysis of handheld GPR measurements, comparing bi-static and multi-static system performance using singular value decomposition (SVD). The results obtained confirmed the conclusions of a previous study conducted by Watson [2], which states that multi-static systems achieve greater subsurface information, for a landmine detection application using full wave inversion (FWI). The study in [2] also conducted a 2D FWI to verify the superior performance of multi-static arrays over a bi-static configuration, for handheld GPR. It concludes that the size of the multi-static array, or number of antenna elements, is insignificant. Furthermore, the acquisition system under test was simplistic and radio propagation properties such as the antenna cross-coupling, radiation pattern and geometry are neglected. Additionally, only a flat, homogenous domain was simulated for the 2D FWI analysis.

Nevertheless, a 2D FWI study is insufficient, as the landmine detection domain is a 3D, heterogeneous domain and a 3D analysis is required to produce a more realistic evaluation of the performance of bi-static and multi-static handheld GPR systems for the demining challenge. Here, an empirical FWI numerical analysis is reported for a 3D domain, which considers a homogenous as well as a heterogeneous or cluttered domain, using a derivative-free optimization algorithm. The bi-static and multi-static system performance are evaluated by comparing the estimated subsurface parameters obtained in each case with synthetic GPR data parameters. All the radio propagation effects are included with simulated experiments conducted in the CST STUDIO SUITE 3D electromagnetic (EM) environment. Finally, a novel method to determine an improved initial parameter set for the GPR FWI optimization, that is closer to the global minimum, is proposed. This is achieved by generating a database of prior forward problem solutions. The database is used with each measured GPR data to determine the least L2 norm objective function, whose parameters are considered as the initial parameter set for the actual FWI optimization.

*Corresponding Author: Suki Dauda Sule, University of Hull, +447443143897
Email: s.d.sule@2008.hull.ac.uk

2. FWI for Multi-static Handheld GPR

2.1. Gradient Based Method

The GPR FWI problem for landmine detection may be posed as a regularised least squares (LSQ) non-linear optimisation problem given by [2]

$$\mathbf{X}_{im} = \arg\min_{\mathbf{x}} \frac{1}{2}\left\|GPR(\mathbf{X}) - \mathbf{d}\right\|_2^2 + \lambda T(\mathbf{X}) \qquad (1)$$

where $\mathbf{X}_{im}$ is a vector of geometric and electrical parameters describing the ground, $\mathbf{d}$ is the GPR measured data, $GPR(\mathbf{X})$ is a forward model that returns the GPR measurement that would be made for a subsurface with parameter vector $\mathbf{X}$. The GPR inverse problem is ill-posed as arbitrarily large changes in $\mathbf{X}$ can have negligible effect on the error $\left\|GPR(\mathbf{X}) - \mathbf{d}\right\|_2^2$. The regularisation term

is required to control the size of components of $\mathbf{X}$ with little or no effects on GPR data. The function $T(\mathbf{X})$ introduces a penalty based on the size of these components and is a way to introduce prior information. For Tikhonov regularisation $T(\mathbf{X})=\|\mathbf{\Gamma X}\|^2$ and often $\mathbf{\Gamma}$ is chosen to be the identity matrix. The regularisation or Tikhonov factor $\lambda \geq 0$ is often adjusted dynamically during the iterative optimisation process to control convergence. Optimisation is posed as a LSQ problem as the error functional uses the L2 norm.

The GPR forward model is non-linear and so (1) is often solved by iterative linearization. Watson used an iterative, quasi Newton method called the limited memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) nonlinear optimisation algorithm [3]. The solution requires a calculation of the gradients of the LSQ error function $\|GPR(\mathbf{X})-\mathbf{d}\|_2^2$. Due to the special form of the LSQ error function, these derivatives can be directly related to the derivatives of the forward model.

2.2. Derivative-free Method

The derivative-free method for solving the FWI problem is used because it is computationally expensive to estimate the derivatives required for derivative based methods. Derivative-free methods require less computation per iteration and are suitable for a limited number of variables, but may require more iterations. A range of non-derivative nonlinear optimization algorithms may be used to solve the FWI optimisation problem when gradients do not exist or are expensive to compute. Local direct search methods may be used when there are a small number of variables and the objective function is computationally cheap to evaluate. The Nelder-Mead simplex algorithm [4] is the most cited, robust and efficient of local direct search methods. The algorithm fundamentally relies on an initial number of points that create a simplex i.e. a set of points spanning a volume in the variable dimensionality considered. The objective function is evaluated at the vertices of the simplex during each iteration to determine the highest objective value, which is used to redefine another vertex that produces a new simplex. Additional new points are produced by moving the vertex with the highest objective value through a series of transformations using the centroid of the associated simplex that include reflection, expansion, internal and external contractions [5]. These steps are iterated until convergence is achieved. Convergence can occur for non-smooth objective functions even when the second derivative of the function is unobtainable [6]. The Nelder Mead simplex algorithm may be applied to the FWI optimisation problem, to estimate the uncertainty in parameter sensitivity of the handheld GPR bi-static and multi-static systems. The Nelder Mead Simplex algorithm is embedded in the CST STUDIO SUITE environment and so the 3D GPR forward model may be integrated with the optimization process on the same platform.

3. Modelling and Simulation

3.1. The GPR System Models and Experiments

The GPR system models are developed with linear antenna arrays positioned with dipoles at a fixed distance above the ground surface. The configurations are for a bi-static system, and multi-static system with two and four receive elements, driven in a single input multiple output (SIMO) sequence. The antennas are end-fed (coaxial) dipole antennas designed and optimised using the Antenna Magus software for a centre frequency of 2.5 GHz and frequency range of 1.75-3.25 GHz. Assuming the same offset of 20 cm used by Watson, an antenna element spacing of 5 cm was used in all configurations. Antenna cross-coupling losses are significant, given this spacing which is less than a wavelength. The antennas are placed initially at a height of approximately 3.76 cm above the ground surface. The transmitting element is at the extreme left end of the array and the time series measured at each receive antenna are concatenated into a single data vector **d**. The target object is placed under the centre of the array, based on the assumption that a metal detector (MD) has located a conducting part of the device.

To reduce the computational cost, the ground size for this study is reduced to a 31 cm by 29 cm box with a depth of 9cm. The subsurface parameters are the relative permittivity $\varepsilon_r^{soil}=2.53$ and loss tangent $\tan\sigma=0.0036$. The target is a single AP mine which is modelled as a typical plastic cylinder with relative permittivity $\varepsilon_r^{mine}=2.8$ (typical US M14 mine). The diameter and height are 7cm and 6cm respectively, closer to a Colombian military MN-MAP-1 mine [7]. The mine also contains a tetryl charge (US M14) with relative permittivity $\varepsilon_r^{tetryl}=2.163\ \left(1g/cm^3\right)$ and an air void of free space relative permittivity of one. Instead of ABC boundary conditions (used by Watson), perfectly matched layer (PML) boundary conditions are applied on all faces of the box (ground) for all models in this study with added space above the antennas on the top of box to simulate the antenna to antenna and antenna to ground propagation. ABC boundary conditions have been studied longer and are generally easier to implement, but the PML region achieves less boundary reflections [8].

For a heterogeneous ground (cluttered domain), we introduce a rough surface to the flat ground (box) surface. The ground surface roughness height is modelled in the form of a Gaussian distribution [9]. Therefore, the box planar surface is modelled with depressions and protrusions to simulate a ground surface with a random height, assuming a Gaussian distribution, normally distributed white noise and a mean value of zero [10]. Additionally, subsurface clutter sources are modelled as several 3D rectangular blocks, grouped into two clusters, with each set having a different relative permittivity. One has a relative permittivity, $\varepsilon_r=5$ and the other with a relative permittivity, $\varepsilon_r=4$. An example bi-static dipole system and target subspace is shown in Figure 1. A LSQ FWI regularised optimisation problem, as in (1), is the test problem. The goal value to be evaluated is the sum squared difference of the total simulated GPR data (time-series) and the total forward model data (time-series), which is given by $\|GPR_{meas}-GPR(\mathbf{X})\|^2$. Watson performed a similar study for dipole antennas in a simplified 2D, homogenous domain only. Here both 3D homogenous and heterogeneous domains are investigated for the FWI analysis.

The optimization is set to a maximum of 20 iterations, due to computational constraints. The hexahedral mesh used to produce

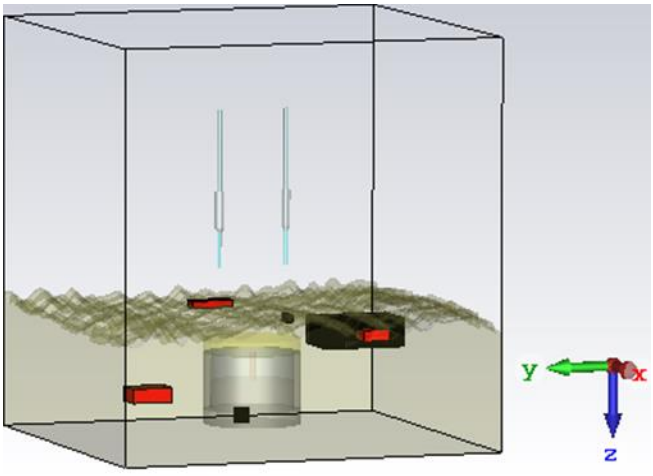


Figure 1: Bi-static dipole system for a heterogeneous ground

synthetic GPR measurements is different from the mesh used in the forward model within the FWI objective function. This avoids the spuriously good results that can be the inverse crime artefact [11] due to using the same mesh for both operations.

The parameter vector for the synthetic GPR data for the homogenous ground is given by

$$GPR_{initial} = [\varepsilon^T, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s] = [2.163, 0.0036, 1, 2.8, 2.53] \quad (2)$$

where ε^T = relative permittivity of tetryl charge

$\tan\sigma$ = loss tangent

ε_0 = relative permittivity of free space

ε_r^m = relative permittivity of plastic mine

ε_r^s = relative permittivity of dry sandy soil

Whereas the parameter vector for the synthetic GPR data for the heterogeneous ground is given by

$$GPR_{initial} = [\varepsilon^T, \varepsilon^{C1}, \varepsilon^{C2}, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s] = [2.163, 3.75, 6, 0.0036, 1, 2.8, 2.53] \quad (3)$$

where ε^{C1} = relative permittivity of first clutter source

ε^{C2} = relative permittivity of second clutter source

ε_0 = relative permittivity of free space

ε_r^m = relative permittivity of plastic mine

ε_r^s = relative permittivity of dry sandy soil

The initial subsurface parameter vector for the homogenous ground FWI is given by

$$GPR_{initial} = [\varepsilon^T, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s] = [2.363, 0.0046, 1.2, 3.01, 2.23] \quad (4)$$

Whereas the initial vector for the heterogeneous ground FWI is given by

$$GPR_{initial} = [\varepsilon^T, \varepsilon^{C1}, \varepsilon^{C2}, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s] = [2.363, 5, 4, 0.0046, 1.2, 3.01, 2.23] \quad (5)$$

3.2. FWI Numerical Analysis Results

To estimate the performance of the different antenna systems, we compare and consider the estimated parameter values approached by the FWI solutions for both models with their respective synthetic GPR data true parameter values. The error in the subsurface parameters for each system in the different ground conditions is estimated by determining the total sum squared difference between the GPR data vector parameters and the FWI solution vector parameters. The total estimated parameter errors are indicated in Tables 1 and 2 for the homogenous ground and heterogeneous ground respectively.

The 3D FWI data shows that for both ground domains, the conclusion that multi-static systems can achieve more subsurface information is achieved as the smallest subsurface error in both cases is recorded with a multi-static system. At this stage, we can confirm that the multi-static system performs better than the bi-static system in general. However, the subsurface parameter error is not monotonic for the number of antenna elements whereas it was found to be linear for a 2D FWI analysis. These results based on a 3D analysis are more realistic as the scattering on the soil surface and cylindrical mine introduce more degrees of freedom and complexity than a 2D numerical analysis.

For the homogeneous soil, the 4 receiver (RX) system FWI achieves a better performance than the two other systems though the difference between the bi-static system and 2 RX multi-static system is marginal. However, in the more realistic heterogeneous domain the 4 RX multi-static system yields the largest parameter uncertainty. The scattering and clutter signals are observed to be larger in the heterogeneous domain than the homogenous domain based on the larger subsurface parameter error and uncertainty figures that are estimated in the former. This is due to greater air-ground reflections from the rough ground surface and scattering from the buried clutter sources. In this case the clutter signal is much larger than the mine signal. Nevertheless, for both soils, particularly the heterogeneous soil, the parameterisation does not sufficiently describe the clutter and so the optimisation easily converges to the wrong solution. Some method is required to reduce the effects of the clutter signal. Therefore, better imaging with multi-static systems for a real GPR system and demining operation based on FWI is predicated on an optimised antenna design as well as clutter reduction.

Conversely, this analysis also shows that the multi-static system achieves only a small improvement over the bi-static system. A FWI optimization with actual GPR measurements or field evaluation data is expected to exhibit greater complexity and hence the improvement may only be marginal. This study has been limited to synthetic data and future studies may include further

validation of the results obtained with the use of measured GPR data.

Table 1: Summary of GPR and FWI solution parameter values for homogenous ground

Subsurface Parameters	GPR Data Parameter Values	FWI Solution Estimated Parameter Values		
		Bi-static	Multi-static (2RXs)	Multi-static (4RXs)
Charge relative permittivity	2.163	2.292	2.268	2.3
Loss tangent	0.0036	0.0026	0.0034	0.0069
Air void relative permittivity	1.0	1.49	1.5	1.45
Mine relative permittivity	2.8	3.069	3.185	3.032
Soil relative permittivity	2.53	2.2	2.2	2.6
Estimated parameter error	**-**	**1.219**	**1.320**	**0.892**

Table 2: Summary of GPR and FWI solution parameter values for heterogeneous ground

Subsurface Parameters	GPR Data Parameter Values	FWI Solution Estimated Parameter Values		
		Bi-static	Multi-static (2RXs)	Multi-static (4RXs)
Charge relative permittivity	2.163	2.4	2.388	2.314
Clutter1 relative permittivity	3.75	2.25	2.25	2.25
Clutter2 relative permittivity	6.0	7.28	6.67	8.0
Loss tangent	0.0036	0.0058	0.0066	0.0068
Air void relative permittivity	1.0	1.35	1.29	1.39
Mine relative permittivity	2.8	3.04	2.96	3.16
Soil relative permittivity	2.53	2.27	2.33	2.2
Estimated parameter error	**-**	**3.869**	**3.048**	**4.734**

4. Improvement of the FWI Initial Parameter Set

The steepest descent and direct search methods for iterative nonlinear optimisation all require an initial parameter set which is updated iteratively according to the chosen method. The GPR FWI problem is also a local minimisation problem that is nonlinear as well as ill-posed. A good initial parameter set that is as close as possible to the true solution is desirable to ensure convergence to the global minimum and less computational expense. A more general approach would be to employ global optimization techniques prior to the local optimization. However, for the GPR problem this would be computationally prohibitive. We propose a compromise approach that involves solving the forward problem for several parameter sets within the bounded local domain or search space and using an L2 norm objective with the true solution or measured data, as the goal value. The forward model meshing or grid would be set to the lowest tolerance level as only a coarse analysis is required to avoid a high computational expense. A higher tolerance is set for the forward model grid for the actual optimization.

A database of the forward model solutions for a specific operation is generated and can be evaluated with data from each GPR measurement. The initial parameter vector for the FWI solution is chosen by interpolation from the database of parameter vectors and measured time-series. Hence a single database is generated during a single campaign but can be used repeatedly for GPR data from the same source environment or location. The database campaign could be done during the training phase of a demining operation prior to the actual clearance operation. The database can be generated for any chosen number of parameter combinations and sample space or bounded conditions. More forward model solutions would be expected to increase the probability of a better initial estimate of the parameter set. For this experiment, the 4 RX dipole system GPR model for a heterogeneous domain is utilised. Eleven forward model solutions are generated by arbitrarily choosing eleven different parameter sets for values between a minimum and maximum bound. This yields a database of eleven sets of A-scan data and the L2 norm objective value of each of these data for any measured GPR is determined. The parameter set for the time-series (A-scan) that achieves the lowest objective function value would be considered as the closest to the true solution or global minimum and selected as the initial optimization parameter set. Figure 2 presents the objective function value for all eleven forward problem solutions.

It can be determined from Figure 2 that the lowest objective function value is obtained at simulation run eight which corresponds to a value of 0.001708. The parameter set for this forward model measurement, GPR_F is given by

$$\begin{aligned} GPR_F &= [\,\varepsilon^T, \varepsilon^{C1}, \varepsilon^{C2}, \tan\sigma, \varepsilon^0, \varepsilon_r^m, \varepsilon_r^s\,] \\ &= [2.25, 5.75, 6.50, 0.0061, 1.35, 3.05, 2.55] \end{aligned} \quad (6)$$

Therefore, this parameter set given in (6) is selected as the initial parameter set for the iterative FWI solution, for the heterogeneous domain under test. The FWI optimization solution for this parameter set is then obtained for 20 iterations, due to computational constraints, and compared directly with the FWI solution for the initial parameter set in (3). The result is shown in Figure 3.

The comparison of the FWI solution for the original initial parameter set and the one derived from the database generation (Figure 3) shows that the latter does not achieve an improvement in the accuracy of convergence as the absolute error is marginally less than the original solution. However, the result does show that the database generated initial parameter FWI solution is closer to the true solution. This could potentially lead to a more efficient optimization using a more powerful algorithm. Derivative based methods would benefit from this improvement and achieve convergence with less computational expense and fewer iterations as a local minimum would be realised more efficiently. The procedure achieves the primary goal of improving the initial parameter estimation or guess. The CST integrated Nelder-Mead optimisation does not allow the user to specify the entire simplex and objective function values. If this had been the case, the initial 10 iterations exploring the simplex could have been entirely

avoided by selecting all the simplex points from the database. The performance of gradient based algorithms are expected to benefit from the database generated improved initial parameter set.

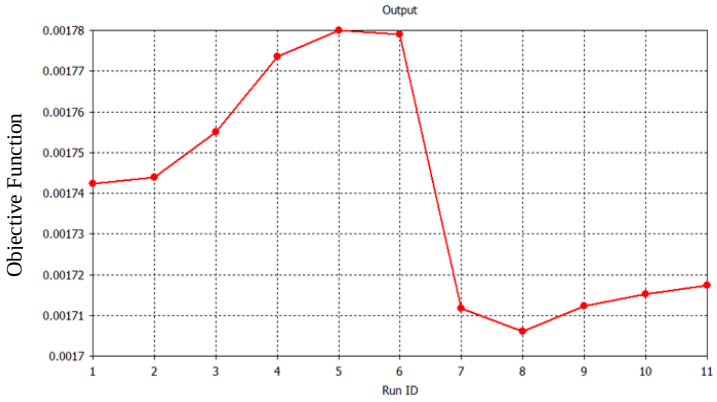


Figure 2: Objective function values for eleven forward problem solutions

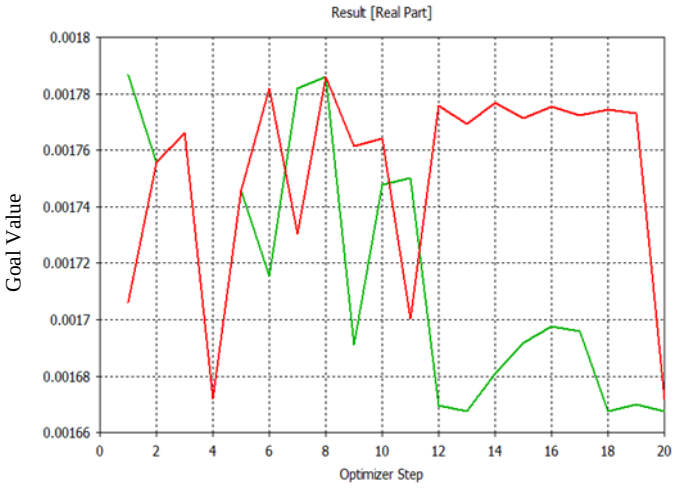


Figure 3: FWI solution result for 4 RX dipole a. original initial parameter set (GREEN) versus database selected parameter set (RED)

5. Conclusion

An empirical study has been undertaken to compare 3D FWI imaging using multi-static and bi-static systems in homogeneous and heterogeneous media. The results verify the possibility of multi-static systems to achieve greater subsurface parameter sensitivity and hence reliability of target detection than bi-static systems. In 2D, all multi-static systems outperformed the bi-static systems. However, a more realistic 3D analysis shows that the improvement in performance with increasing numbers of antennas is not simple or monotonic due to cross-coupling and antenna patterns. This underlines the need for optimisation of the antenna system configuration and size (number of elements), to achieve better performance than a bi-static system. Additionally, the effect of clutter significantly limits the accuracy of parameter estimation. Finally, a novel procedure has been proposed to determine the initial parameter vector for the FWI solution which yields an initial forward problem solution that is closer to the true GPR data solution. The procedure requires numerous forward problem solutions stored prior to a deeming campaign but has the potential to significantly reduce the computational expense of the FWI as well as the accuracy for an ideal local minimisation.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to acknowledge the Petroleum Technology Development Fund (PTDF), Abuja, Nigeria.

References

[1] Sule, S.D. and Paulson, K.S., "A comparison of bistatic and multistatic handheld ground penetrating radar (GPR) antenna performance for landmine detection", IEEE Radar Conference (RadarConf), 2017, pp. 1211-1215. **DOI:** 10.1109/RADAR.2017.7944389

[2] Watson, F. and Lionheart, W., "SVD analysis of GPR full-wave inversion", 15th International Conference on Ground Penetrating Radar (GPR), 2014, IEEE, pp. 484-490. **DOI:** 10.1109/ICGPR.2014.6970472

[3] Nocedal, J. and Wright, S., Numerical optimization. Springer Science & Business Media, New York, 2006.

[4] Nelder, J.A. and Mead, R., "A simplex method for function minimization", The computer journal, **7**(4), pp. 308-313, 1965. https://doi.org/10.1093/comjnl/7.4.308

[5] Rios, L.M. and Sahinidis, N.V., "Derivative-free optimization: a review of algorithms and comparison of software implementations", Journal of Global Optimization, pp. 1-47, 2013. doi:10.1007/s10898-012-9951-y

[6] Mckinnon, K.I., "Convergence of the Nelder--Mead Simplex Method to a Nonstationary Point", SIAM Journal on Optimization, **9**(1), pp. 148-158, 1998. https://doi.org/10.1137/S1052623496303482

[7] Lopera, O. and Milisavljevic, N., "Prediction of the effects of soil and target properties on the antipersonnel landmine detection performance of ground-penetrating radar: A Colombian case study", Journal of Applied Geophysics, **63**(1), pp. 13-23. https://doi.org/10.1016/j.jappgeo.2007.02.002

[8] Nataf, F., "Absorbing boundary conditions and perfectly matched layers in wave propagation problems, Direct and Inverse problems in Wave Propagation and Applications", 14, de Gruyter, Radon Ser. Comput. Appl. Math., pp.219-231, 2013, 978-3-11-028228-3. https://hal.archives-ouvertes.fr/hal-00799759

[9] Tajdini, M.M., Gonzalez-Valdes, B., Martinez-Lorenzo, J.A., Morgenthaler, A.W. and Rappaport, C.M., "Efficient 3D forward modeling of GPR scattering from rough ground", IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting, pp. 1686-1687, 2015. **DOI:** 10.1109/APS.2015.7305232

[10] Daniels, D.J., "The impact of antenna design on short range radar performance", IEEE Conference on Antenna Measurements & Applications (CAMA), pp. 1-4, 2014. **DOI:** 10.1109/CAMA.2014.7003338

[11] Kaipio, J. and Somersalo, E., "Statistical inverse problems: discretization, model reduction and inverse crimes", Journal of Computational and Applied Mathematics, **198**(2), pp. 493-504, 2007. https://doi.org/10.1016/j.cam.2005.09.027

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 130-134 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Modelling of the resistance heating of the moving molybdenum sheet

Miroslav Pavelek[*,1], Michal Frivaldsky[1], Pavol Spanik[1], Tibor donič[2]

[1]*Department of mechatronics and electronics, University of Zilina, 01026, Slovakia*

[2]*Research Centre of University of Zilina, University of Zilina,01026, Slovakia*

ARTICLE INFO

Article history:
Received: 02 January, 2018
Accepted: 17 January, 2018
Online: 30 January, 2018

Keywords :
Resistance heating
Molybdenum sheet
Finite element method

ABSTRACT

This article describes the modeling procedure of a direct resistive heat model of the molybdenum sheet and its simulation with electrodes made from various materials. The main characteristic of the proposed model is its dynamic behavior, because model considers the movement of the molybdenum sheet. The simulation results are mutually compared and the optimal material parameters are selected. The development of the model, in which the movement of molybdenum sheet is considered is used for determination of optimal movement speed of the molybdenum sheet in order to achieve requested temperature. The presented model can be also used for determination of optimal electrode materials and its geometrical properties/shape.

1. Introduction

This paper is an extension of work originally presented in Electrical Power engineering 2017 [1]. This article discusses the usage of finite element method for the development of a resistive heating model of the molybdenum sheet. The model is designed for the needs of the development of the equipment for the heating and shaping the molybdenum sheets into molds. The mentioned molds are used as containers for horizontal crystallization of the sapphire single crystal. Molybdenum is one of the few materials suitable for this purpose in a view of the long-term high-temperature load of the mold. The mold is exposed to the temperature gradient of up to 100 °C./ mm at the maximum temperature of 2150 °C. in the longitudinal direction of the container, where in one part is the melt and in the adjacent part is the corundum as an already growth product, i.e. single sapphire crystal. Based on the extreme conditions during the process, it is crucial to not decrease the mechanical abilities during the shaping process of the mold. For that purpose, the structural changes and possibilities of avoiding them by properly set pre-heating respectively heating/cooling system of Mo sheets need to be predicted. Currently, the molds are made of molybdenum sheet having a thickness of 0,5 mm, made by powder metallurgy technology, i.e. by plastic deformation and sintering procedure. Specifically, the molybdenum sheet Grade M1 corresponding to the American standard ASTM B386 or GB 3877, where the chemical composition is approximately the same. [2-5]

[*]Corresponding Author: Miroslav Pavelek, University of Zilina
Email: miroslav.pavelek@fel.uniza.sk

2. Basic simulation model

The simulation model is developed in COMSOL environment as a 3D model with the options for reconfiguration the mutual position of electrodes, the electrode material and the pressure on the electrodes. The model is composed of the molybdenum sheet and two electrodes. The molybdenum sheet is modeled as a block (domain) with a wanted width(a), depth(b) and thickness(c) and electrodes are modeled as cylinders with a wanted diameter(d), height(h) and distance(x) between them. The geometry (Fig.1.) is then complemented by physic setup. The Multiphysics "Joule heating (JH)" is used for the model. This Multiphysics is formed by the connection of "Electrical circuit (EC)" physic and "Heat transfer in solid (HTS)" physic. The physic "Electrical circuit" is used to set up the value of input current trough top boundary of one electrode and grounding through the top boundary of the other electrode. Within the EC module, the required contact pressure applied to the electrodes and the roughness of the contact between the electrode and Mo sheet can be adjusted. The HTS is used to set up parameters as initial temperature of the system and the method in which the temperature is transferred to the surrounding area. The proper function of the model is achieved by setting the spherical domain with a diameter of one meter, which is filled with air. All simulations are the set as time domain simulation, so we can determine the influence of studied materials on the speed of the heating of the Mo sheet. [6-9]

As it has been already mentioned, the model is reconfigurable, but for means of this paper the following geometrical parameters are set: Parameters of molybdenum sheet (based on manufactured molybdenum sheets [10,11]): **1000/300/1 (a/b/c) mm,** parameters

of electrodes: **20/50 (d/h) mm**, the electrodes distance: **100 (x) mm**. The electrodes are pressed to the Mo sheet with a pressure of **0.3 MPa.** For the electrode's material study, three classes of materials based on the electrical conductivity of the electrodes are defined (lower than Mo, Mo and higher than Mo). The electro-thermal parameters of the basic model are defined in TABLE I. The model needs to be particularly accurate in the molybdenum sheet domain. For that purpose, the temperature dependencies of electro-thermal parameters of molybdenum are implemented, specified by the manufacturer "PLANSEE" (Fig.2-4.).

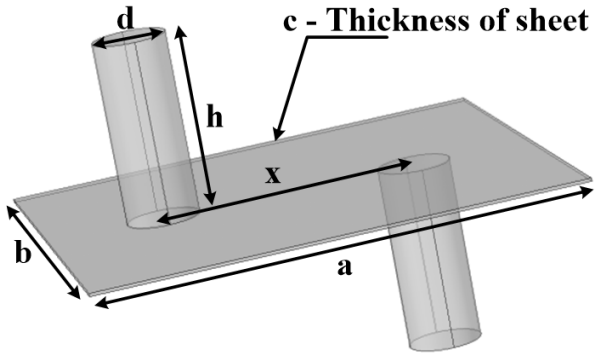


Fig. 1. Representation and definition of Mo sheet model

Table 1: Electro-thermal parameters used in simulations

Material	Parameter				
	Electrical cond. [S/m]	*Relative perm. [-]*	*Thermal cond. [W/(mK)]*	*Specific heat capacity [J/(kgK)]*	*Density [kg/m^3]*
Molybdenum (20 °C) (293.15 K)	17.9e6	1	142	254	10280
Class I. Lower El. Con.	$\sigma_{Mo}/10$	1	400	385	8700
Class II. Equivalent El. Con.	σ_{Mo}	1	400	385	8700
Class III. Higher El. Con.	$\sigma_{Mo}*100$	1	400	385	8700

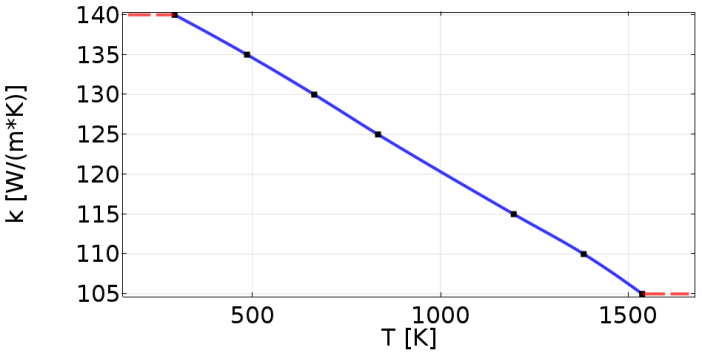


Fig. 2. Thermal dependence of thermal conductivity of molybdenum used in simulations

The mentioned JH module of the COMSOL environment uses well-known equations for modeling a heat transfer in solid materials (1).

$$\rho C_P \frac{\partial T}{\partial t} - \nabla \cdot (k \nabla T) = Q \qquad (1)$$

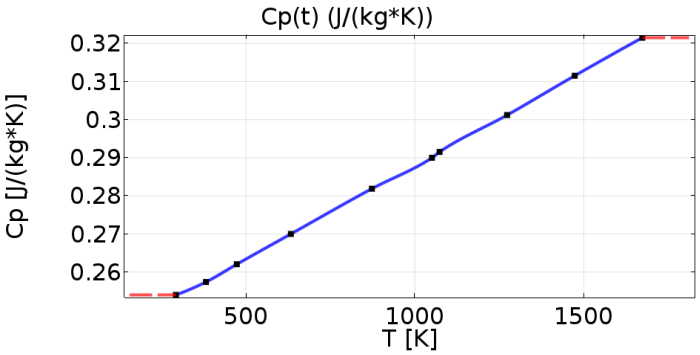


Fig. 3. Thermal dependence of specific heat capacity of molybdenum used in simulations

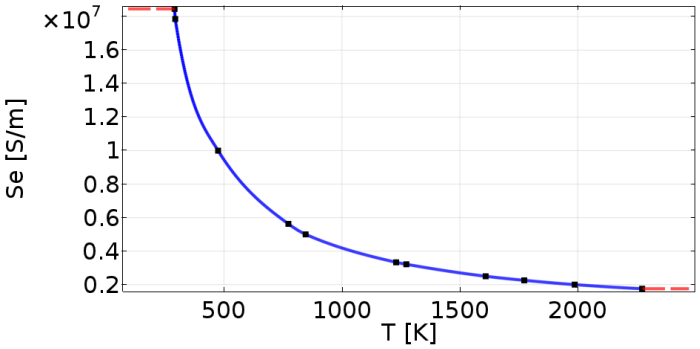


Fig. 4. Thermal dependence of electrical conductivity of molybdenum used in simulations

Where "ρ" is the density of the modeled material, "C_P" is specific heat capacity of the material, "k" is thermal conductivity of the material and "Q" is a heat source.

In this case, the heat source "Q" is determined by Joule-Lenz law of resistive loses in material structure due to the current flow through this structure.

3. Simulation results I.

Based on [6] the wanted temperature of Mo sheet is at a level between 1000 K to 1200 K. At those levels, the Mo sheet can be easily bendable. The simulation is done for before mentioned electrode's material classes (TABLE I.) not for specific material. The input current of the system is selected within the interval from 1000 A to 1400 A. The results are compared in a way of temperature distribution within Mo sheet domain and speed of the heating process. Fig.5. shows actual simulation model with geometrical parameters defined in the previous paragraph.

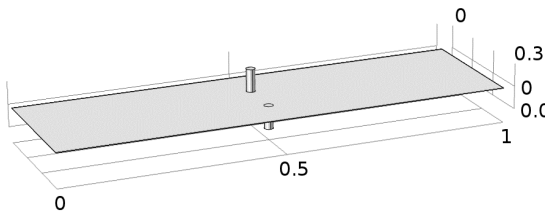


Fig. 5. Model of Mo heating with given geometrical parameters

The simulation is done with electrodes placed in the middle of the longer side of the sheet while the distance between electrodes is set to 100 mm. The following results are displayed as the temperature in the point in the middle of the sheet's volume and temperature distribution in the cutline between electrodes (middle of the Mo sheet) as can be seen at Fig.6.

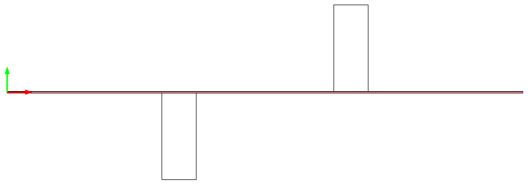


Fig. 6. Line for result extraction (Temperature distribution)

Fig.7. shows mentioned results for the first class of electrode's material, which has lower electrical conductivity than Molybdenum. The wanted temperature of 1000 K (Fig.7. up) was achieved in 75 s for lowest current (1000 A) and in 4 s for highest simulated current (1400 A). The wanted temperature interval is marked as black dashed lines in Fig.7,8,9 (up). The second part of the graph (Fig.7.middle) shows temperature distribution (based on Fig.6.) for different current values, while the time when the desired temperature value was achieved (1000 K) is considered. The last part of the graph shows temperature distribution within whole Mo sheet with the chosen current flow. Fig.8. and Fig.9. have the same purpose as Fig.7., but other material types are considered. For second electrode's material (Se=17.9e6 S/m) can be seen that wanted temperature was achieved in 95 s for current value of 1400 A.

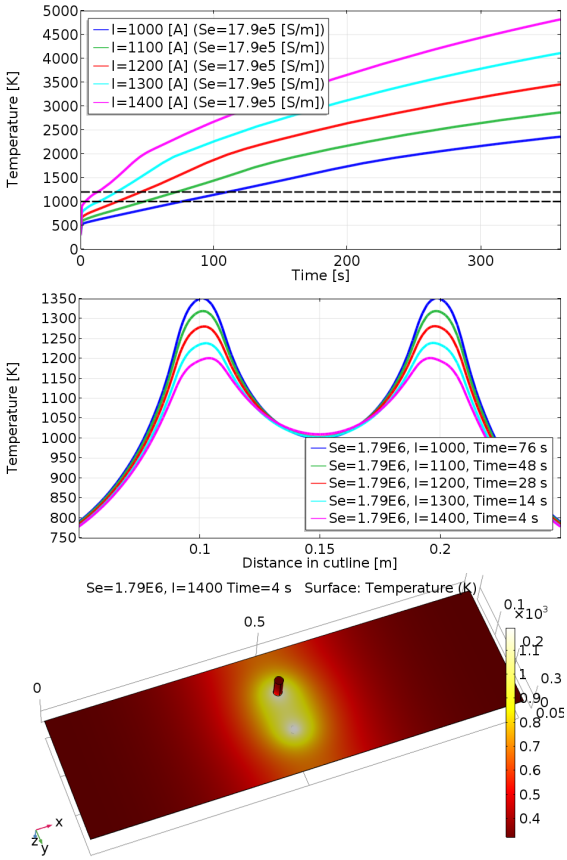


Fig. 7. Results for first electrodes material type (Se = 17.9e5 [S/m])

In case of third material the heating process was much slower (the major part of heat is produced through the direct resistive heating, while in first two cases there was also considerable influence of indirect resistive heating), wanted temperature was achieved in 164s for current 1400 A and it can be said that minimal current needed to achieve wanted temperature in given time is 1200 A.

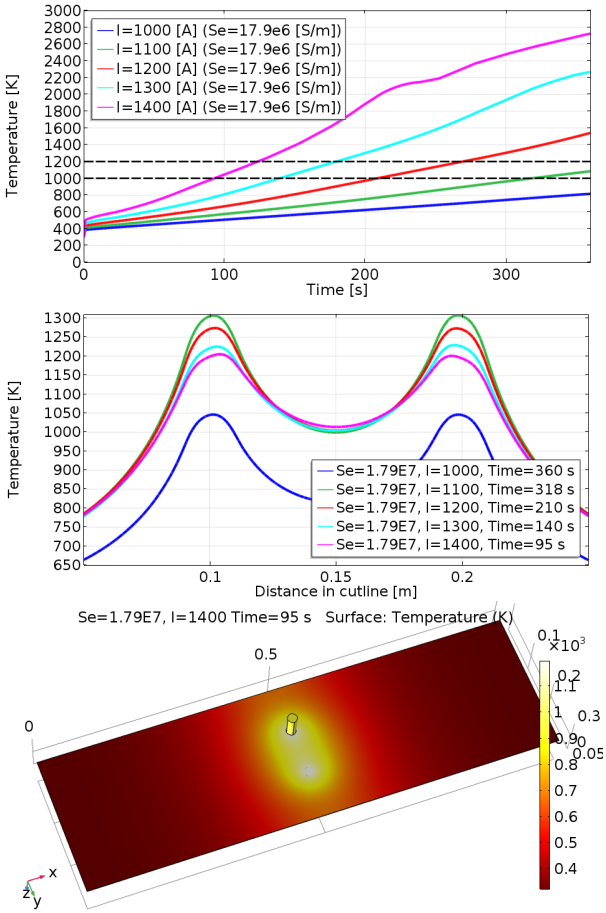


Fig. 8. Results for second electrede's material type (Se=17.9e6 [S/m])

4. Model of sliding molybdenum plate

For simulation of sliding molybdenum sheet the same settings of physics and geometry are used as within the basic model (previous part), but for the shift of the Mo sheet a special script in Matlab language is developed, which basically simulates the heating in a loop with the different initial position of electrodes. For the required temperature distribution, which is different for each step of the simulation, the molybdenum sheet domain is divided into subdomains (blocks) (Fig.10.). Within individual subdomains, the average temperature is defined (volumetrically) from previous simulation step and next simulation step is simulated with given electrodes shift. Each domain is within script defined as a 3D object with given length, width and height. These geometrical parameters are computed from geometrical parameters of whole Mo plate and from a wanted number of the subdomain in every direction of Mo plate. Mathematically, this model is described as the basic simulation model with equation 1 and by Joul-Lenz law for heat source of the modeled physic.

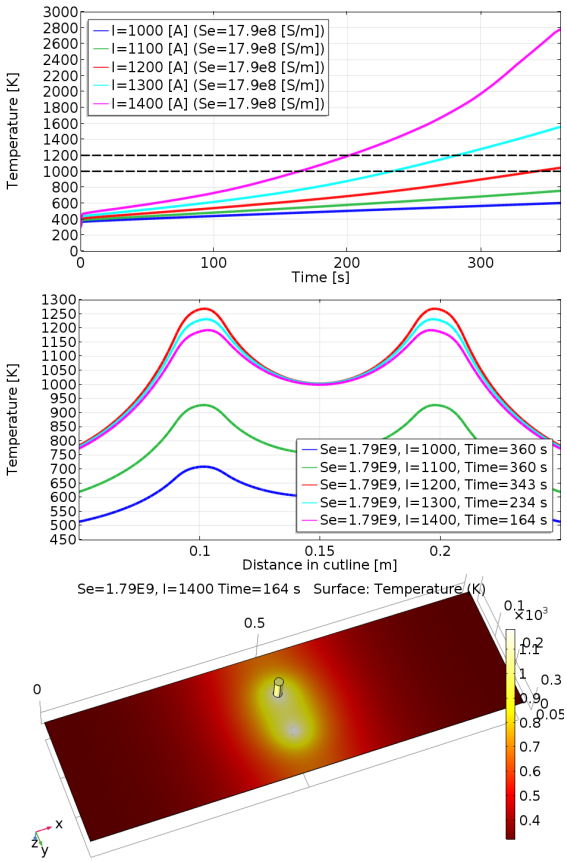


Fig. 9. Results for third electrode's material type (Se=17.9e8 [S/m])

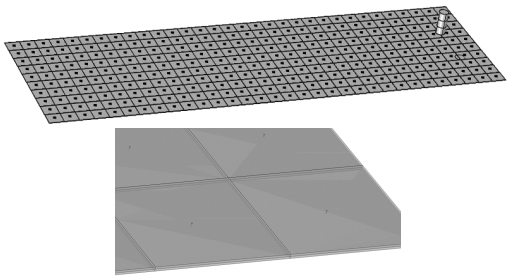


Fig. 10. Model of Mo heating with subdomains

5. Simulation results II.

The simulation is done for the second class of materials with electrical conductivity value close to Mo sheet, in which the temperature distribution was most uniform. Based on the optimal speed of the heating process and distribution of heat within Mo sheet, the input current of 1400 A is applied. It needs to be mentioned that this part of the model was created purely for verification of the proposed modeling process and thus, it does not consist any actual comparison with measurement. For that reason, a quite large (120 mm) electrodes shift is chosen for simulation. Model of the Mo sheet is composed of (32x10x2) subdomains and it is simulated in 5 steps with total electrodes shift of 600 mm in 100 s. The results shown in fig. 11 demonstrates the functionality of the proposed model (chapter 4), while each step of simulation is defined two times. The initial time (0 s) is representing initial temperature setting within each subdomain with new electrode position. The secondary time (20 s) is representing end time for each step and for the temperature distribution needed what is needed for next step's initial condition. Based on the results of the individual simulation steps, it can be said that the modeling process is suitable for this application and may be applied also in other cases/studies.

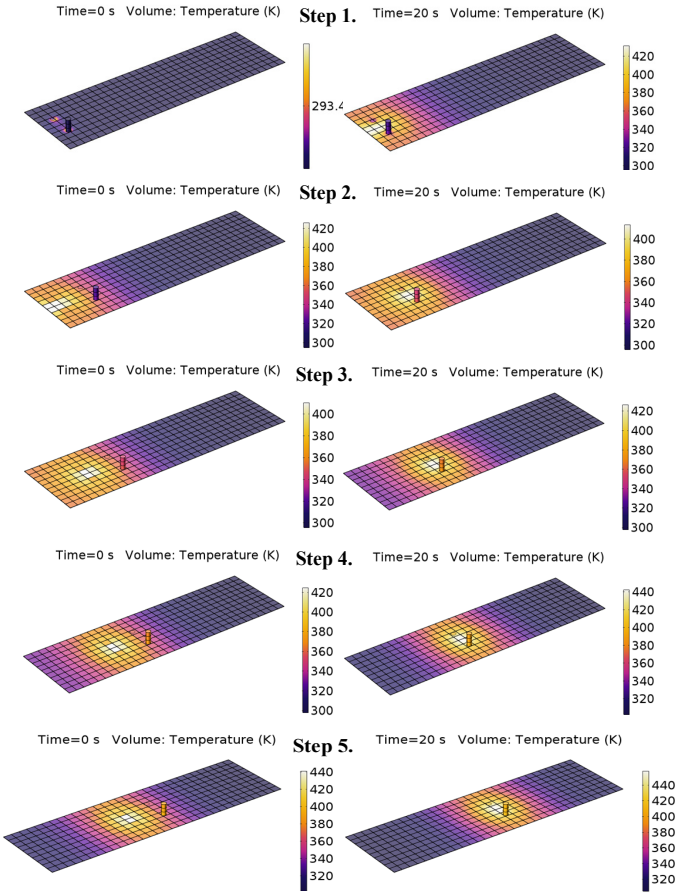


Fig. 11. Simulation results for given parameters

6. Conclusion

In this paper, the design of thermal finite element simulation model of Mo sheet was described. The main aim was to target exact modeling of thermal field in the structure of Mo sheet, while multiple material properties of heating electrodes where considered because resistive heating is considered within the proposed paper.

The presented model had the possibility of the reconfiguration of the electrode material, the electrodes distance, the electrode geometry as well as Mo sheet geometry. Also, Multiphysics is considered, whereby various values of electric current have been applied during experiments. For these purposes, several types of simulation models were developed. The main problem during the investigation is the proper definition of the material properties of Mo sheets. Each manufacturer has different material composition. Therefore, the simulation model has to have the possibility of reconfiguration in order to meet experimental results, which have been made within previous researchers. Consequently, the presented model shall serve for optimization procedures, because experimental investigation mostly acts as time-consuming. Based on investigated configuration it can be claimed that there are possibilities of improvement in speed of heating (point-heating)

as well as possibility of pre-heating (line-heating) by using electrodes from different materials.

Second approach o presented paper was focused on the design of the model, which will consider a dynamic change of the heated place within Mo sheet, i.e. translation move was applied. Various software products have limitations regarding the dynamic movement of the investigated sample, therefore special script in MATLAB environment was developed. With the use of this approach, each iteration of the simulation accepts previous results, which are repeatedly implemented within the computation solution.

Given proposal of the solution was required due to future works, which will be focused on the design of multi-physics simulation model of Mo sheet heating and molding system, where high-validity simulation models are expected to be used namely for further investigations of mechanical parameters and other molybdenum restrictions.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors wish to thank Slovak grant agency APVV for the project no. 0396-15 - Research of perspective high-frequency converter systems with GaN technology and for the project no. 14-0284 - Study of Utility Properties of Formed Molybdenum Sheets Applicable for Horizontal Crystallization of Sapphire Single Crystal.

References

[1] M. Pavelek, M. Frivaldsky, P. Spanik, T. Donic, “Electrothermal model of resistance heating of molybdenum sheet” in EPE 2017, Kouty nad Desnou 2017. https://doi.org/10.1109/EPE.2017.7967262

[2] L. Hargaš, M. Hrianka, J. Lakatos, D. Koniar, “Heat Fields Modeling and Verification of Electronic Parts of Mechatronics Systems” Metalurgija **49**(2), 268 – 272, 2012. https://doi.org/10.5755/j01.eee.123.7.2371

[3] P.Karban, F.Mach, I.Doležel, “Hard-couple model of local direct resistance heating of thin sheets” Journal of Computational and Applied Mathematics **236**(18), 4725-4731, 2012. https://doi.org/10.1016/j.cam.2012.02.036

[4] F.Dughiero, M.Forzan, C.Pozza, E. Sieni, “A translational couple electromagnetic and thermal innovative model for induction welding of tubes” IEEE Transactions on Magnetics **48**(2), 483-486, 2012. https://doi.org/10.1109/TMAG.2011.2174972

[5] K.Mori, S. Maki, Y.Tanaka, “Warm and Hot Stamping of ultra-high tensile strenght sheets using resistance heating” CIRP Annals-Manufacturing Technology **54**, 209-212, 2005. https://doi.org/10.1016/S0007-8506(07)60085-7

[6] S. Maki, Y. Harada, K.I. Mori, H. Makino, “Application of resistance heating technique to mushy state forming of aluminium alloy” Journal of materials processing technology **125**, 477 – 482, 2002.

[7] J. Cuntala, A. Kondelova, O. Hock, M. Pridala, “Electro-Thermal Modeling of Power LED Using COMSOL Environment” in ELEKTRO 11TH INTERNATIONAL CONFERENCE, Slovakia, 2016.

[8] M. Hrianka, L. Hargaš, J. Lakatš, D. Koniar, “Modeling, Simulation, and Verification of Heat Transfer in Power Transistor Cooler” Metalurgija 2 **49**(), 283 – 287, 2012.

[9] K.Mori, S. Maki, Y.Tanaka, “Warm and Hot Stamping of ultra-high tensile strenght sheets using resistance heating” CIRP Annals-Manufacturing Technology, **54**(), 209-212, 2005.

[10] Plansee Group, https://www.plansee.com/en/index.html [10.11.2017]

[11] ED FAGAN INC., MACHINING GUIDE, http://www.edfagan.com/litPDF/Machining_Guide_All_Materials.pdf [10.11.2017]

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 135-147 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Algorithms for Technical Integration of Virtual Power Plants into German System Operation

André Richter*, Ines Hauer, Martin Wolter

Otto-von-Guericke University, Institute of Electric Power Systems, Chair Electric Power Systems and Renewable Energy Sources, 39106 Magdeburg, Germany

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 20 December, 2017
Online: 30 January, 2018

Keywords :
Virtual Power Plants
Operational concepts
Power system
Renewable Energy Sources
Optimization
Case study

ABSTRACT

This paper critically evaluates the operational perspective of Virtual Power Plants (VPP) in Germany by analyzing key factors to replace conventional power plants in the future power system. Therefore, its necessity for the secure operation as well as the technical and economic benefits for the German power system are pointed out. The single sections describe in detail how the requirements on technical functions and standardized communication can be reconciled with the increasing challenges on volatile generation. Furthermore, different operation concepts (profit maximization, intra-day schedule loyalty and system services maximization) are described with respect to their mathematical algorithms and their practical feasibility under consideration of given circumstances and future developments. The impact on the power system by the different possible VPP-operational concepts are exemplary pointed out in a case study with use of a medium voltage Cigre-Benchmark Network. The results indicate a high impact on hand-over points by VPP operational concepts.

1. Introduction

This paper is an extension of work originally presented in 14th International Conference on the European Energy Market (EEM) [1]. The goal of the German government to reduce CO2 emissions about 40 % until 2020 leads to a rising energy infeed fluctuation evoked by renewable energy sources in the German power system[2]. Due to a high percentage of volatile power infeed, the share of gross electricity generation in 2016 increased to 12.3 % from wind, 7.0 % from bio power plants and 5.9 % from photovoltaic (PV) [3]. The future share of renewable energy sources should rise up much more according to the German government. 45 % of the renewable power was installed in medium and low voltage level [4], so that a large percentage of uncontrolled generation is located in this area. Therefore, the electricity system requires intelligent power networks to secure a stable balance between generation and demand [5]. One part of this intelligent power network is a Virtual Power Plant (VPP) that provides centralized control for aggregated units. The future system shall provide advanced functionalities similar to real power plants aiming on economically sensible generation and increased reliability [6, 7]. Therefore, the individual functionalities of decentralized units are bundled up using suitable communication and control systems to entirely replace conventional power plants [7]. However, bundling up single functionalities of single generation unit's does not directly secure similar operation concepts to conventional power plants.

Therefore, this paper deals with the question how VPP can be integrated into the current German power system exploiting their advantageous characteristics like centralized control of small volatile renewable generation units. Included in this, current available technical VPP properties as well as available and economic reasonable operation concepts are checked under consideration of current market circumstances. For example, so far there is no operation concept for VPP that supports the elimination of critical situations in TSO and DSO voltage levels.

[6 – 8] present some investigations already dealt with corresponding optimization methods for operational concepts, but none of them addresses the overall system requirements in its entirety. For that reason, Figure 1 illustrates the system design of a VPP as well as the interaction of market and power system participants and components. All units are remotely accessible through a central energy management and control center, which calculates the optimal schedule by taking into account information from the electricity market and forecast data of the individual generation and load units.

The overall flexibility of decentralized power units could be improved by applying the concept of VPP's [6, 7, 9] but the German electricity market together with applicable regulations still

*Corresponding Author: André Richter, Email: andre.richter@ovgu.de

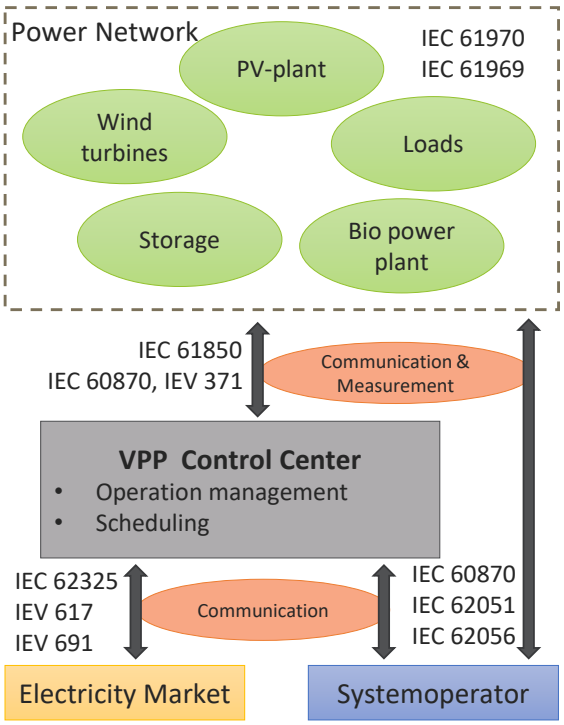


Figure 1. Components of the VPP and influencing factors

not provide the required incentives for a sustainable operation and consequently for a large-scale roll-out. Based on the identification of basic requirements, this paper points out different necessary presupposes for the operation of VPP's similar to conventional power plants. This includes different aspects like communication, generation and demand forecast as well as the German electricity market behavior and costs (*c*) for generation (**Error! Reference source not found.**).

Table 1: Costs of renewable generation without investment costs

	Variable costs in € /kWh	**Annual fix costs in € /kW**
Wind turbines [10]	0.0241 – 0.0268	56 – 61
PV plants [11, 12]	0.0000	13.00 – 35.00
Biopower plants [12]	0.0325	175.00

Furthermore, these presupposes are necessary to realize the new VPP operational concepts for taking part in network congestions elimination or control reserve contribution as well as intra-day schedule loyalty concerning the day-ahead planned generation unit operation. Therefore, not only the benefit for VPP owners and direct marketers is identified, but mainly the operational concpets are mathematically described. The paper presents an ongoing case study, where operational concepts are tested and evaluated according to the impact in a medium voltage CIGRE benchmark system. Thereby, the basic preconditions for the operation of Virtual Power Plants according to the German Market system are considered. Finally, the results and conclusion will indicate the high impact of small centralized controlled generation units as well as use cases where only small incentives are needed for the realization of the system supporting technical concpets. This paper differs strongly from the original paper in [1] by an extended definition of market restriction, in detail concerning the control reserve market and by presenting defined operational concepts and algorithms.

2. Basic Requirements of VPP's

2.1. Communication

Communication is a basic requirement for the operation of VPP as well of intelligent power networks in general. Therefore, it is necessary that generation units, controllable loads and all players offering power flow flexibility and aiming on benefits from variable tariffs can use a standardized and reliable communication system [5]. Based on this, there is a need for a standardized communication language and easy data exchange between all players [13]. Table II lists typical standards for communication in electric power systems, which can be applied in German power system. Furthermore all these standards are necessary for the operation of VPP and the communication with measurement systems, market and the system operator.

Table 2: Communication standards for intelligent power networks [14, 15]

Standard content	**Standard**
Telecontrol	IEC 60870, IEV 371
Communication systems in substations	IEC 61850
Interface for distribution management	IEC 61968
Energy market communication	IEC 62325
Data exchange with metering equipment / Electricity Metering	IEC 62051 /IEC 62056
Application in energy management system	IEC 61970 / 61969

Table II shows that various communication standards are applied to VPPs [16, 17]. According to [17] the IEC 61850 represents one of the most important communication standards for the operation and control of VPP with the integrated intelligent networks and distributed units. The innovative standard enables a secure and reliable communication and control using intelligent measurement systems like Phasor Measurement Units (PMUs) and control units like Remote Terminal Units (RTUs) [17 – 19]. Its necessity is based on the requirement to standardize data-models and furthermore to describe communication mechanisms and system management aspects. This standard is an extension to IEC 60870 resulting from adapted power system requirements. IEC 61850 has the advantage of defining data units through logical nodes with defined data structures (common data classes) [20]. The adressed protocol enables the control center to have direct data access and direct control to any physical unit of the VPP in the power network. Furthermore, the application of this standard combined with direct communication supplies the VPP operator with information about the physical location so that he is able to separate physical units according to accounting areas for instance.

2.2. Forecast-oriented power unit scheduling

For setting up an optimal power unit schedule, the VPP operator needs to distinguish the individual characteristics of the available resources. This information is not only necessary for an

optimal market participation, but also for enabling the system operator to make day-ahead and intra-day forecasts as well as to know the maximum controllable power for critical situations. The demand forecast for high numbers of small consumers can be realized using standardized load profiles. Bigger customers, like industry companies with bilateral contracts, require separate measurement systems in order to meet the requirements for providing dedicated demand side potential. In contrast, a larger percentage of units in the VPP are generation units and especially volatile ones like PV plants and wind turbines. Consequently, an exact forecast for those units is indispensable from the operator‘s point of view, e.g. for scheduling the secondary control reserve [21].

The forecast for wind and PV generation can be realized with different methods presented in Figure 2 and Figure 3. Thereby, the volatile factors are wind speed and solar radiation [19].

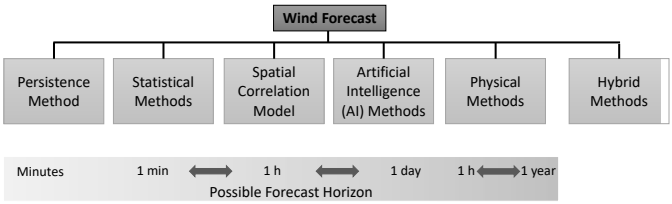


Figure 2. Overview of wind forecast methods and corresponding mathematical algorithms, [21 – 23]

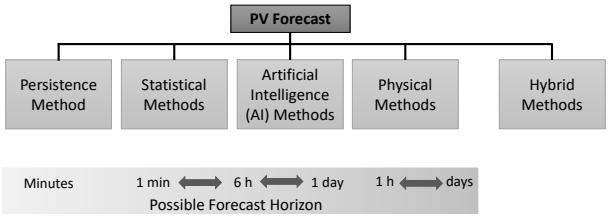


Figure 3. Overview of PV forecast methods and corresponding mathematical algorithms [25]

Wind forecast methods can be divided in six groups depending on its corresponding mathematical algorithms. Four methods (Persistence-, Statistical-, Artificial Intelligence method and Spatial Correlation model) are based on current wind measurement and statistical data as well as on learning approaches to predict wind generation. Depending on the applied method, the forecast horizon varies between a few minutes up to one day. To achieve an increase in accuracy and time horizon, it is necessary to use a physical or hybrid method. Both methods are based on meteorological and climate data and use the numerical weather prediction (NWP) for that purpose, but under the constraint of high computational effort. Regarding the methods and the forecast horizon the best accuracy (mean absolute error) varies between 5 % and 7 % for one to two hours forecast and increases for day-ahead forecast [21 - 24]. According to Figure 3, three of the shown methods (Persistence-, Statistical- and Artificial Intelligence Method) process current solar radiation and statistical data. This way utilizable data can be provided for a forecast horizon of a few minutes until one day. However, similar to wind forecast the accuracy rapidly decreases with the forecast horizon. Compared to wind forecast, PV forecast depends even more on the availability of information from weather stations and satellite data to detect clouds, high fog and other phenomena that effect solar radiation. Therefore, the physical method and the hybrid methods use NWP and Cloud Imagery for prediction with subsequent data post processing [25].

In consequence to the described uncertainties in generation forecast, the VPP operator has to deal with different impacts. In case of direct market participation, he has to expect sanctions from the power system operator or to participate in the Intra-Day market accepting possible high costs [26]. Besides using storage systems, there is a solution by using existing flexibilities from wind turbines pv, biomass plants and other technical solutions as in **Error! Reference source not found.**. Those technologies can provide power reserve if they are not operated at their maximum power output limit. However, since storage systems are still expensive and an appropriate market model involving positive and negative power capacity of wind and bio generation is not provided, more profit can be achieved by operating the units at its maximum power output level.

2.3. German Electricity Market

In this section, the German electricity market will be characterized focusing on VPP to identify possible operation behavior which influence system operation. Based on this, there are three possibilities to sell energy from renewable energy sources. The first possibility is to receive fixed payment, which is determined based on the German Renewable Energy Act (EEG). The other two possibilities are to participate in the Energy-Only-Market and to sell control reserves at the balancing market [26].

The electricity is traded in Germany either at the energy exchange or bilateral at the Over-the-Counter market. The European Energy Exchange (EEX) in Leipzig is the biggest market place for electricity produced in Germany. Electricity on the EEX can be traded either on the futures market or on the spot market (EPEX SPOT in Paris). The main difference between these markets is the electricity delivery time. Thus, the future markets provides contracts for long-term supply of electricity up to six years in the future [26]. The spot market is used as a trading platform for the short-term supply of electricity within 1-2 days, namely Day-Ahead and Intra-Day market [26, 28]. Thus, the day-ahead market is used to trade the electricity, that will be generated and consumed the next day. Because of the fact, that the day-ahead market is based on forecasted profiles (generation and consumption), an Intra-Day market is necessary to balance the forecasted errors. The contracts on the Spot market are mostly carried out physically, that means the physical electricity delivery to the fixed date [29]. According to (§11 EEG 2014), a system operator is obligated to transmit the renewable electricity and to sell it at the exchange. The operators of renewable power plants receive a fixed remuneration, which is supported by the EEG-apportionment. As a result, the renewable energy sources are always first at the merit-order, independently of the operation costs. Nuclear and lignite power plants follow, since they are usually used to cover the base load due to technical and economic advantages. Hard coal, gas and oil power plants are usually more expensive. Thus, renewable electricity has a great influence on the electricity price, because of its prior position. The operator of renewable energy sources can freely choose between a fixed EEG remuneration and the direct marketing at the beginning of each calendar year, §20 (1) EEG 2014. In addition, all systems with an installed power of more than 100 kW, which are set to operation from 1 January 2016, must be marketed directly (§ 21 EEG 2017). The benefit of direct trading in contrast to the fixed remuneration is that trader get the chance of higher profits by receiving market earnings plus a market premium. The market premium is lower

than the fixed remuneration but another governmental incentive to get a higher percentage of renewables in the market system. However, there are some technical limitations which restrict the operation of VPP at the market and in the power system. These are for example the lowest bid of 0.1 MW to participate in the Day-ahead and in the continous Intra-Day market. Furthermore, there are finacial requirements to participate in the spot market, in Germany [30]. The participation fee for Day-ahead market is about 10000 €/ a with additional 0.04 €/MWh (Day-Ahead auction) and 0.08 €/MWh (intra-day auction). Aditionally, there is a one-time participation fee to get into the spot market of about 25000 € for Day-Ahead market. These are costs that have to be covered by an optimal operational concept or new regulatory incentives.

The control reserve market in Germany is characterized by the fact that the German transmission system operator are responsible to keep the power system stable and the rated frequency of 50 Hz. There are 3 types of control reserves (primary control reserve, secondary control reserve as well as minute reserve). These types are differentiated according to activation speed and duration. The procurement of these control reserves is organized in a tender auction with participation of power plant operators and customers. One main fact is that pooling is still permitted to reach the market entry boundaries [31]. Main characteristics of the three control reserves are explained subsequently according to [32]:

Primary control reserve (PCR):

- full activation of PCR within 30 seconds
- incident period that have to be covered 0 <t <15 min
- automatic activation
- symmetrical (positive and negative PCR is not separated) bid of minimum +/- 1 MW
- the tender submission period is one week
- fixed power price

Secondary control reserve (SCR):

- energy balance of the control area and frequency control
- immediate automatic activation by the concerned TSO
- complete activation within 5 minutes
- automatic activation based on Merit-Order-List
- separated bid for negative and positive SCR is possible
- separation of SCR positive and negative in high tariff (HT, on Monday till Friday, 8 am til 8 pm) and low tariff (LT, weekend, national holidays and 8 pm til 8 am)
- minimum bid of 5 MW and an increment of 1 MW
- the tender submission period is one week
- fixed power price and energy price on call

Minute reserve (MR):

- automatic activation is based on a Merit-Order-List
- complete activation within 15 minutes
- incident period that have to be covered are t > 15 min to 4 quarter hours or up to severalhours in case several incidents
- separated bid for negative and positive MR is possible
- positive and negative MR, always with six 4-hour time slices
- minimum bid of 5 MW
- the tender submission period is one day
- fixed power price and energy price on call

Finally, the market regulatories and EEG incentives for the operation of single generation units are nowadays much higher, instead of VPP-operation with centralized controlled of small generation units. But there are ongoing changes in the market system which enable VPP for offering more system supporting functionalities. In [33] a change in the control reserve market system is already anounced. For instance, SCR and MR are intended for a daily (calendar) tender submission period with 4-hour time slices. Furthermore, exceptional rules that allow for minimum bids below 5 MW are introduced.

In conclusion, there are still no high financial and regulatory incentives to operate VPP in a way, which serves both areas – market and technical operation – at the moment. Because of this physical delivery of electricity remains uncertain with respect to a future power system including a higher share of renewables. Consequently, it will be indispensable to use the VPP benefits of generation unit aggregation and its aggregated central controlled flexibilities to have minimum system effecting controllable power. Therefore, some assumptions are made for the ongoing analyzes for a medium-term future VPP.

3. VPP -Operational concepts

One of the main challenges of VPP operators is the setup of the internal merit order, that defines which power unit shall be used to satisfy a certain power demand or infeed. This has to be done under consideration of type of generation unit, its technical capability and regulatory as well as economical restrictions. However, right now VPP operators maximize their profits by offering as much energy as possible in the markets aiming on the highest possible profit. As one result, VPP functionalities remain unused due to missing monetary or regulatory incentives, in particular the centralized control of aggregated small generation units. This causes a problem for the system operator: VPP operators sell energy based on the market premium at every point of time, even in case of congestions in the DSO or TSO network.

Therefore, this paper offers two main system supporting operational concepts and additonally an economical concept (see Figure 4). Each concept is explained in detail in the ongoing subsections 3.2 and 3.3.

1) Economical oriented (green): porift maximization, by optimal energy selling on the electricity market.
2) Technical oriented (orange): intra-day schedule loyalty and system service maximization including centriole reserve maximization and redispatch maximization.

3.1. Mathematical optimization problem

The operational concepts can be mathematically involved and expressed in terms of a maximization or minimization problem (objective function) with a given number of restricted goods (constraints). Therefore, this paper presents options to solve the optimization problems in MATLAB for day-ahead scheduling and intra-day schedule loyalty in intervals of 15 minutes. The development of schedules in day-ahead and intra-day planning rely on an optimization problem with an application-specific objective function. The restrictions are defined for technical characteristics of the system, electricity market restrictions and for current as well as predicted system states. Linear optimization method based on simplex algorithm is applied to solve the optimization problem. Linear optimization either minimizes or maximizes a linear

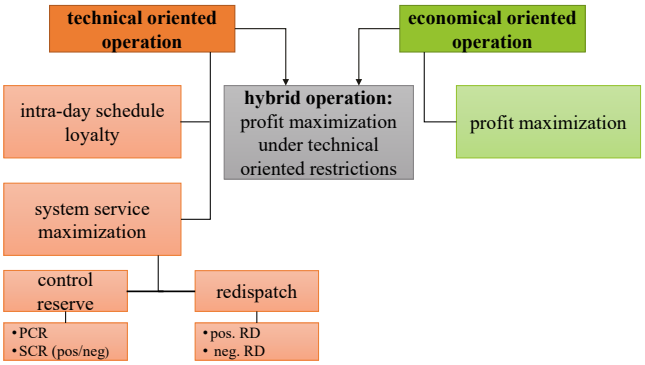


Figure 4. VPP operational concepts

objective function, taking into account linear constraints. The linear optimization contains only continuous variables, whereby benefitery short calculation times are realized. From the linear optimization, the mixed integer linear optimization can be distinguished. Mixed integer linear optimization (MILP) is the basis of the optimization problem for day-ahead and intraday optimization implemented in this paper. Unlike linear optimization, variables can be either continuous or integer or binary. Especially with binary variables it is possible to model logical conditions. The degree of detail of the optimization can be increased, but this is accompanied by an increase in the computational effort for the determination of the optimal solution. [34] The solver intlinprog is used to implement the scheduling problem in MATLAB. Firstly, the objective function is defined and implemented according profit maximization. Ongoing the different technical and economical upper and lower bounds as well as linear constraints were defined and implemented. Finally, the integer constraints have to be implemented. The solver finds a first less constraint starting solution by using linear optimization without integer constraints. The starting solution serves as a starting point for new solutions for the method branch and bound. The method branch and bound takes all constraints and inequaltities into account by generating subproblems and finally finds the best solution, if this solution is within a defined tolerance gap according to the other solutions.

3.2. Economical oriented operation

The main objective of a power plant operator and thus a VPP operator is always to maximize its profits, regardless of the system situation. The regulatory framework and the optimization skills always determine whether or in what amount the operator generates profit. Consequently, the composition of the VPP has a significant impact on potential profits. In the analysis of this concept for resource planning, an optimization based on today's market data for control reserve markets ($p_{\mathrm{PCR/\,SCR/\,MR}}$ / $P_{\mathrm{PCR/\,SCR/\,MR}}$), Day-Ahead auction (p_{auction} / $P_{\mathrm{wind/pv/bio,auc}}$) and continuous Intra-Day market ($p_{\mathrm{continous,IH,T}}$ / $P_{\mathrm{wind/pv/bio,continous,IH,T}}$) is made. The analysis assumes that both a fixed EEG allowance and a market premium for market participation through EEG plants are not paid out. Aim of this assumption is to analyze the VPP operation under perfect market conditions without any subsidies. This leads to the objective function where profit (p) should be maximized by optimal day-ahead generation unit schedule(P_{wind}, P_{pv} and P_{bio}). The optimal distribution of power to different selling options is limited to the forecast of renewables ($P_{\mathrm{wind/pv,\,forecast}}$), the installed power, various constraints and other secondary conditions which are described in (2) – (34).

$$\begin{aligned}\max f(p) = {} & p_{\mathrm{wind}}(t)\cdot P_{\mathrm{wind}}(t) + p_{\mathrm{pv}}(t)\cdot P_{\mathrm{pv}}(t) \\ & + p_{\mathrm{bio}}(t)\cdot P_{\mathrm{bio}}(t) \\ & + p_{\mathrm{auction}}(t)\cdot P_{\mathrm{wind,auc}}(t) \\ & + p_{\mathrm{auction}}(t)\cdot P_{\mathrm{pv,auc}}(t) \\ & + p_{\mathrm{auction}}(t)\cdot P_{\mathrm{bio,auc}}(t) \\ & + p_{\mathrm{continous,IH,T}}(t)\cdot P_{\mathrm{wind,continous,IH,T}}(t) \\ & + p_{\mathrm{continous,IH,T}}(t)\cdot P_{\mathrm{pv,continous,IH,T}}(t) \\ & + p_{\mathrm{continous,IH,T}}(t)\cdot P_{\mathrm{bio,continous,IH,T}}(t) \\ & + p_{\mathrm{PCR}}(t)\cdot P_{\mathrm{PCR}}(t) \\ & + p_{\mathrm{SCR,pos}}(t)\cdot P_{\mathrm{SCR,pos}}(t) \\ & + p_{\mathrm{SCR,neg}}(t)\cdot P_{\mathrm{SCR,neg}}(t) \\ & + p_{\mathrm{MCR,pos}}(t)\cdot P_{\mathrm{MCR,pos}}(t) \\ & + p_{\mathrm{MCR,neg}}(t)\cdot P_{\mathrm{MCR,neg}}(t)\end{aligned} \tag{1}$$

During the analyzes several constraints are considered. The main constraints are:

- The profit calculation based on revenue (r) is depending on the market selling option and the costs (c) (energy generation costs, spot-market selling).

$$p_{\mathrm{wind}}(t) = -c_{\mathrm{wind}} \tag{2}$$

$$p_{\mathrm{pv}}(t) = -c_{\mathrm{pv}} \tag{3}$$

$$p_{\mathrm{bio}}(t) = -c_{\mathrm{bio}} \tag{4}$$

$$p_{\mathrm{auction}}(t) = r_{\mathrm{auction}}(t) - c_{\mathrm{auction}} \tag{5}$$

$$p_{\mathrm{continous.IH,T}}(t) = r_{\mathrm{continous.IH,T}}(t) - c_{\mathrm{continous.IH}} \tag{6}$$

- Minimum possible generation for wind, PV and bio generation that can be used for optimization: The factors $f_{\mathrm{wind,min}}$ $f_{\mathrm{pv,min}}$ and $f_{\mathrm{bio,min}}$ set the lower bounds of minimum generation output. They define pre-process depending generation output especially for bio generation. The minimum for wind and PV is zero by default.

$$f_{\mathrm{wind,min}} \cdot P_{\mathrm{wind,forecast}}(t) \le P_{\mathrm{wind}}(t) \tag{7}$$

$$f_{\mathrm{pv,min}} \cdot P_{\mathrm{pv,forecast}}(t) \le P_{\mathrm{pv}}(t) \tag{8}$$

$$f_{\mathrm{bio,min}} \cdot P_{\mathrm{bio,inst}} \le P_{\mathrm{bio}}(t) \tag{9}$$

- Bio generation power ramping: The maximum rate of change of power ($f_{\Delta P\mathrm{bio,pos}}$, $f_{\Delta P\mathrm{bio,neg}}$) is limited, comparable with conventional power plants.

increase of power for t = 1:

$$P_{\text{bio}}(t) - f_{\text{bio,Start}} \cdot P_{\text{bio,inst}} \leq f_{\Delta P\text{bio,pos}} \cdot P_{\text{bio,inst}} \tag{10}$$

increase of power for t ≥ 2:

$$P_{\text{bio}}(t) - P_{\text{bio}}(t-1) \leq f_{\Delta P\text{bio,pos}} \cdot P_{\text{bio,inst}} \tag{11}$$

decrease of power for t =1:

$$f_{\text{bio,Start}} \cdot P_{\text{bio,inst}} - P_{\text{bio}}(t) \leq f_{\Delta P\text{bio,neg}} \cdot P_{\text{bio,inst}} \tag{12}$$

decrease of power for t ≥ 2:

$$P_{\text{bio}}(t-1) - P_{\text{bio}}(t) - \leq f_{\Delta P\text{bio,neg}} \cdot P_{\text{bio,inst}} \tag{13}$$

- Limitation of provision of control reserve (CR) by bio generation: Not every bio power plant is able to fulfill the technical requirements of PCR provision. Therefore, the proportion of bio generation units that can provide PCR can be limited with the factor $f_{\text{Bio,PCR}}$. Furthermore, the contribution of PCR and SCR has to be defined in constraints according to the technical requirements that PCR have to be provided in at least 30 s and SCR in at least 5 min. $f_{\text{d}P\text{bio,pos}}$ and $f_{\text{d}P\text{bio,neg}}$ describe the maximum possible rate of change of power in one minute so that the constraints for PCR and SCR are described as follows:

PCR contribution within 30 s, positiv and negative:

$$\text{d}P_{\text{bio,PCR,pos}}(t) \leq \frac{f_{\text{d}P\text{bio,pos}}}{2} \cdot f_{\text{bio,PCR}} \cdot P_{\text{bio,inst}} \tag{14}$$

$$\text{d}P_{\text{bio,PCR,neg}}(t) \leq \frac{f_{\text{d}P\text{bio,neg}}}{2} \cdot f_{\text{bio,PCR}} \cdot P_{\text{bio,inst}} \tag{15}$$

SCR contribution within 5 min, positiv and negative:

$$\text{d}P_{\text{bio,SCR,pos}}(t) \leq 5 \cdot f_{\text{d}P\text{bio,Pos}} \cdot P_{\text{bio,inst}} \tag{16}$$

$$\text{d}P_{\text{bio,SCR,neg}}(t) \leq 5 \cdot f_{\text{d}P\text{bio,neg}} \cdot P_{\text{bio,inst}} \tag{17}$$

- Maximum possible control reserve. The provision of control reserve is limited to the requirements described in section 2.3. These are the continuous availability of CR according to the defined periods (PCR-continuous over the day, SCR-LT, SCR-HT, MR-4 hour intervals) and under consideration of wind and PV forecast and the limited contribution of bio generation, respectively.

PCR positive and negative:

$$\begin{aligned} P_{\text{PCR}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + \text{d}P_{\text{bio,PCR,pos}}(t) \end{aligned} \tag{18}$$

$$\begin{aligned} P_{\text{PCR}}(t) \leq\ & P_{\text{wind}}(t) - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + \text{d}P_{\text{bio,PCR,neg}}(t) \end{aligned} \tag{19}$$

SCR positive and negative:

$$\begin{aligned} P_{\text{SCR,pos,HT}}(t) + P_{\text{SCR,pos,NT}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + \text{d}P_{\text{bio,SCR,pos}}(t) \end{aligned} \tag{20}$$

$$\begin{aligned} P_{\text{SCR,neg,HT}}(t) + P_{\text{SCR,neg,NT}}(t) \leq\ & P_{\text{wind}}(t) \\ & - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) \\ & - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + dP_{\text{bio,SCR,neg}}(t) \end{aligned} \tag{21}$$

MR positive and negative:

$$\begin{aligned} P_{\text{MCR}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + P_{\text{bio,inst}} - P_{\text{bio}}(t) \end{aligned} \tag{22}$$

$$\begin{aligned} P_{\text{MCR}}(t) \leq\ & P_{\text{wind}}(t) - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + P_{\text{bio}}(t) - f_{\text{bio,min}} \cdot P_{\text{bio,inst}} \end{aligned} \tag{23}$$

- Avoidance of multiple use of power for control reserve provision: This constraint defines an upper limit for the sum of CR not exceeding the amount of available power. This is divided into positive and negative CR.

$$\begin{aligned} CR_{\text{pos}}(t) = & P_{\text{PCR}}(t) + P_{\text{SCR,pos,HT}}(t) \\ & + P_{\text{SCR,pos,NT}}(t) + P_{\text{MR,pos,4 h-slice}}(t) \end{aligned} \tag{24}$$

$$\begin{aligned} CR_{\text{pos}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + \text{d}P_{\text{bio,PCR,pos}}(t) \end{aligned} \tag{25}$$

$$\begin{aligned} CR_{\text{pos}}(t) \leq\ & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + P_{\text{bio,inst}} - P_{\text{bio}}(t) \end{aligned} \tag{26}$$

$$\begin{aligned} CR_{\text{neg}}(t) = & P_{\text{PCR}}(t) + P_{\text{SCR,neg,HT}}(t) \\ & + P_{\text{SCR,neg,NT}}(t) + P_{\text{MR,neg,4 h-slice}}(t) \end{aligned} \tag{27}$$

$$\begin{aligned} CR_{\text{neg}}(t) \leq\ & P_{\text{wind}}(t) - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + \text{d}P_{\text{bio,SCR,neg}}(t) \end{aligned} \tag{28}$$

$$\begin{aligned} CR_{\text{neg}}(t) \leq\ & P_{\text{wind}}(t) - f_{\text{wind,min}} \cdot P_{\text{wind,forecast}}(t) \\ & + P_{\text{pv}}(t) - f_{\text{pv,min}} \cdot P_{\text{pv,forecast}}(t) \\ & + P_{\text{bio}}(t) - f_{\text{bio,min}} \cdot P_{\text{bio,inst}} \end{aligned} \tag{29}$$

- Minimum bid and increment for control reserve provision, according to the market restrictions: The control reserve amount can be zero or or an integer

multiple of the increments, depending on size of minimum bid and minimum increment. The restriction is presented for PCR with analog constraints for SCR and MR. The factor $y(t)$ is a binary variable for the proof, if enough power is available for CR provision.

$$P_{\mathrm{PCR}}(t) - y_{\mathrm{PCR}}(t) \cdot P_{\mathrm{PCR,min}} \geq 0 \tag{30}$$

- Spot-market restriction: According to the spot-market options regarded in this paper, there are the possibilities to sell the energy in the Day-Ahead auction offered by hour or in the continuous Intra-Day market offered by 15 minutes. Because of the fact that this is a day-ahead planning process, Intra-Day market prices have to be assumed as forecasted values. A further constraint is that the available energy should be sold out under profitable conditions.

$$P_{\mathrm{wind}}(t) = P_{\mathrm{wind,auc}}(t) + P_{\mathrm{wind,continous,IH,T}}(t) \tag{31}$$

$$P_{\mathrm{pv}}(t) = P_{\mathrm{pv,auc}}(t) + P_{\mathrm{pv,continous,IH,T}}(t) \tag{32}$$

$$P_{\mathrm{bio}}(t) = P_{\mathrm{bio,auc}}(t) + P_{\mathrm{bio,continous,IH,T}}(t) \tag{33}$$

$$\begin{aligned} & P_{\mathrm{wind}}(t) + P_{\mathrm{pv}}(t) + P_{\mathrm{bio}}(t) \\ & = P_{\mathrm{wind,auc}}(t) + P_{\mathrm{wind,continous,IH,T}}(t) \\ & + P_{\mathrm{pv,auc}}(t) + P_{\mathrm{pv,continous,IH,T}}(t) \\ & + P_{\mathrm{bio,auc}}(t) + P_{\mathrm{bio,continous,IH,T}}(t) \end{aligned} \tag{34}$$

As a result of the economic optimization schedules are defined for wind, PV and bio generation. All presented constraints consider the figures presented in section 2. This methodology allows for an economically optimal day-ahead schedule, regardless of the size of the VPP (available generation units) and the impact on the electrical power system. In future systems with high penetration of RES, this can lead to unstabelizing effects and to a lack of system services. Therefore, the system itself and system operator need system supporting services to handle exemplary high volatility generation and forecast errors.

3.3. Technical oriented operation

Within this section, various new network-relevant VPP operating concepts are to be examined, which are intended to ensure stable grid operation in the future and also to minimize the effort of operational management for directly and indirectly connected grid operators. These different concepts thus represent the integration of network operator requirements into the operational planning. The new technical concepts meet the request of reliable schedules contributed by VPP and by renewable generation units in general. Furthermore, the presented concepts will enable a maximum of PCR and SCR contribution as well as redispatch. Therefore, the day-ahead maximization function in (1) with its constraints in (2) – (34) are applied. Changed or added objective functions and constraints for optimal day-Ahead scheduling will be presented according to the concepts. The intra-day schedule loyalty represents a further concept which was presented in general in [35].

3.3.1. Intra-day schedule loyalty

The objective of intra-day schedule loyalty is to minimize the schedule deviation, which is mainly based on forecasting errors related to renewable energy source (RES) infeed in the operational planning of the VPP control center. In this case, the power consumption or the power recovery from or into the higher-level network should be reduced, so that the higher-level network operator or the accounting area manager only has to make few or no interventions. If all VPP operators were capable of doing this in the future, day-ahead planning process of system operators can be optimized and network congestions may be avoided.

However, this requires a defined level of flexibility within the VPP and can also lead to curtailments of generation units. The intra-day loyalty operation must be preceded by a day-ahead optimization in which capacities are blocked in order to compensate intra-day schedule deviations. The consumption of control energy should be avoided, if possible. For this purpose, a high fictitious penalty of 1000 €/MW is initially assumed for the balancing energy consumption when positive balancing power ($p_{\mathrm{reBAP,pos}}(t)$) or negative balancing power ($p_{\mathrm{reBAP,pos}}(t)$) is consumed. If the day-ahead schedule can not be met without consumption of balancing energy, balancing energy can be received ($P_{\mathrm{reBAP,pos}}(t)$ and $P_{\mathrm{reBAP,neg}}(t)$). Therefore, the term of the balancing energy represents a measure of the deviation from schedule. Thereby, a main constraint is that a control reserve deviation ($P_{\mathrm{deviation,CR,pos}}(t)$ and $P_{\mathrm{deviation,CR,neg}}(t)$) is prohibited by use of fictitious penalty ($p_{\mathrm{deviation,CR}}$) about 10000 €/MW. The objective function is presented in (35) with the objective to minimize the deviation according to the planned market schedule.

$$\begin{aligned} \max f(p,P) = & \; p_{\mathrm{wind,intraday}}(t) \cdot P_{\mathrm{wind,intraday}}(t) \\ & + p_{\mathrm{pv,intraday}}(t) \cdot P_{\mathrm{pv,intraday}}(t) \\ & + p_{\mathrm{bio,intraday}}(t) \cdot P_{\mathrm{bio,intraday}}(t) \\ & + p_{\mathrm{reBAP,pos}}(t) \cdot P_{\mathrm{reBAP,pos}}(t) \\ & + p_{\mathrm{reBAP,neg}}(t) \cdot P_{\mathrm{reBAP,neg}}(t) \\ & + p_{\mathrm{deviation,CR}} \cdot P_{\mathrm{deviation,CR,pos}}(t) \\ & + p_{\mathrm{deviation,CR}} \cdot P_{\mathrm{deviation,CR,neg}}(t) \end{aligned} \tag{35}$$

This operational concept requires similar constraints, like fictitious penalties for CR deviation and the possible use of balancing power. But the most important point for this operational concept is the flexibility, which is used to fulfill the schedule. Therefore, it is needed to reduce the maximum available power output from bio-generation in the day-ahead planning process by a defined value (e.g. 5 %). This procedure has to be applied for wind and PV forecast as well. The day-ahead forecast has to be reduced by a factor to ensure intra-day flexibilities and intra-day schedule loyalty. The factor has to be optimized depending on forecast precision and VPP generation unit composition. Two further constraints are loyalty of generation units and observance of day-ahead and intra-day schedule. The constraint loyalty of generation units has the aim to consider optimized day-ahead generation unit schedules for wind, PV and bio under the changed intra-day

conditions for renewable infeed ($P_{\text{wind,intraday}}(t)$ and $P_{\text{pv,intraday}}(t)$). If this will not be possible, energy has to be bought according to positive balancing power ($P_{\text{reBAP,pos}}(t)$) or negative balancing power ($P_{\text{reBAP,pos}}(t)$). The constraint observance of day-ahead and intra-day schedule ensure that the schedules of generation units can be changed according to the failure in forecast of renewables. But the sum of generation and the planned amount of cumulated power in each time step for trading on Day-Ahead and Intra-Day market is as much as possible prohibited to change. Both constraints can be expressed as:

$$\begin{aligned} P_{\text{wind}}(t) + P_{\text{pv}}(t) + P_{\text{bio}}(t) = {} & P_{\text{wind,intraday}}(t) \\ & + P_{\text{pv,intraday}}(t) \\ & + P_{\text{bio,intraday}}(t) \\ & + P_{\text{reBAP,pos}}(t) - P_{\text{reBAP,neg}}(t) \end{aligned} \tag{36}$$

The result of this operational concept is a reliable schedule planning which supports system operator in systems with a high amount of RES to keep a stable system and avoid redispatch. The changed generation unit schedule depends on the quality of forecast and reserved capacity for intra-day flexibility and thus indirectly depending on the quality of the forecast.

Furthermore, this concept only needs small monetary or regulatory incentives, because the needed flexibility reserve not leads to high loss in profit. But there is a need of incentives for this operational concept.

3.3.2. Redispatch maximization

The concept of redispatch maximization is a mainly technical triggered concept, which currently has no regulatory or monetary incentives. But in medium-term and future systems there will be an incredible high need for system operators to handle congestions in a system with less controllable conventional power plants. VPP can provide those functions utilizing the concept of centralized control of RES. It has to be assumend that the VPP is either small, so that it feed in a defined network section like in one medium voltage network, or that the schedule for this concept is adaptable for defined network sections. This assumption has to be done to ensure the effectiveness of redispatch in the system. The objective function (37) is focused on the maximization of negative ($P_{\text{RD,neg}}(t)$) or positive ($P_{\text{RD,pos}}(t)$) available power for redispatch. Therefore, the monetary objective function (1) is extended with fictious positive ($p_{\text{RD,pos}}(t)$) and negative ($p_{\text{RD,neg}}(t)$) profit for redispatch. This profit has to be given a high constant value only. A benefit of the methodology of fictious profits to maximize technical needs is that minimum monetary incentives can be identified under given market circumstances, if there is an ambition to create and examine new market concepts for technical needs.

With the integration of redispatch, additional constraints have to be defined. These are the definition of lower bounds for available power, exceeding a defined minimum in (38) – (39) and the calculation of available power for redispatch in (40) – (41).

$$\begin{aligned} \max f(P_{\text{RD}}) = {} & p_{\text{Wind}}(t) \cdot P_{\text{Wind}}(t) + p_{\text{PV}}(t) \cdot P_{\text{PV}}(t) \\ & + p_{\text{Bio}}(t) \cdot P_{\text{Bio}}(t) \\ & + p_{\text{auction}}(t) \cdot P_{\text{Wind,auc}}(t) \\ & + p_{\text{auction}}(t) \cdot P_{\text{PV,auc}}(t) \\ & + p_{\text{auction}}(t) \cdot P_{\text{Bio,auc}}(t) \\ & + p_{\text{continous,IH,T}}(t) \cdot P_{\text{Wind,continous,IH,T}}(t) \\ & + p_{\text{continous,IH,T}}(t) \cdot P_{\text{PV,continous,IH,T}}(t) \\ & + p_{\text{continous,IH,T}}(t) \cdot P_{\text{Bio,continous,IH,T}}(t) \\ & + p_{\text{PCR}}(t) \cdot P_{\text{PCR}}(t) \\ & + p_{\text{SCR,pos}}(t) \cdot P_{\text{SCR,pos}}(t) \\ & + p_{\text{SCR,neg}}(t) \cdot P_{\text{SCR,neg}}(t) \\ & + p_{\text{MCR,pos}}(t) \cdot P_{\text{MCR,pos}}(t) \\ & + p_{\text{MCR,neg}}(t) \cdot P_{\text{MCR,neg}}(t) \\ & + p_{\text{RD,pos}}(t) \cdot P_{\text{RD,pos}}(t) \\ & + p_{\text{RD,neg}}(t) \cdot P_{\text{RD,neg}}(t) \end{aligned} \tag{37}$$

$$P_{\text{RD,pos}}(t) \geq P_{\text{min,RD,pos}}(t) \tag{38}$$

$$P_{\text{RD,neg}}(t) \geq P_{\text{min,RD,neg}}(t) \tag{39}$$

$$\begin{aligned} P_{\text{RD,pos}}(t) = {} & P_{\text{wind,forecast}}(t) - P_{\text{wind}}(t) \\ & + P_{\text{pv,forecast}}(t) - P_{\text{pv}}(t) \\ & + P_{\text{bio,inst}} - P_{\text{bio}}(t) \end{aligned} \tag{40}$$

$$\begin{aligned} P_{\text{RD,neg}}(t) = {} & P_{\text{wind}}(t) + P_{\text{pv}}(t) \\ & + P_{\text{bio}}(t) - f_{\text{bio,min}} \cdot P_{\text{bio,inst}} \end{aligned} \tag{41}$$

3.3.3. Control reserve maximization

Control reserve is one of the most needed products in systems with high volatile generation. CR is important for a global point of view for systems with increased percentage of renewables and decreased number of CR providing conventional power plants. VPP offers the possibility to fulfill market restriction for control reserve market by pooling small decentralized generation units. Depending on the size (amount of installed renewable capacity), VPP can provide only CR products most of the time. Therefore, this optimization maximizes the contribution of positive or negative SCR or the contribution of PCR. The only limiting factors are market restriction concerning minimum bid size and minimum increment, so that CR is maximized first and remaining energy is traded according to profit maximization. This optimization is again based on (1) extended by fictitious high profit for CR. The fictitious profit is chosen very high, depending on the maximization problem (PCR, pos. SCR and neg. SCR).

$$\begin{aligned}\max f(P_{CR}) = {} & p_{Wind}(t)\cdot P_{Wind}(t) + p_{PV}(t)\cdot P_{PV}(t) \\ & + p_{Bio}(t)\cdot P_{Bio}(t) \\ & + p_{auction}(t)\cdot P_{Wind,auc}(t) \\ & + p_{auction}(t)\cdot P_{PV,auc}(t) \\ & + p_{auction}(t)\cdot P_{Bio,auc}(t) \\ & + p_{continous,IH,T}(t)\cdot P_{Wind,continous,IH,T}(t) \\ & + p_{continous,IH,T}(t)\cdot P_{PV,continous,IH,T}(t) \\ & + p_{continous,IH,T}(t)\cdot P_{Bio,continous,IH,T}(t) \\ & + p_{PCR}(t)\cdot P_{PCR}(t) \\ & + p_{SCR,pos}(t)\cdot P_{SCR,pos}(t) \\ & + p_{SCR,neg}(t)\cdot P_{SCR,neg}(t) \\ & + p_{MCR,pos}(t)\cdot P_{MCR,pos}(t) \\ & + p_{MCR,neg}(t)\cdot P_{MCR,neg}(t)\end{aligned} \quad (42)$$

The only constraints which have to be added according to the day-ahead optimization are the definition of lower bounds (P_{min}) for minimum provision of control reserve. Further restrictions are still defined in section 3.2. The constraints are:

$$P_{PCR}(t) \geq P_{min,PCR} \quad (43)$$

$$P_{SCR,pos,HT}(t) \geq P_{min,SCR,pos,HT} \quad (44)$$

$$P_{SCR,pos,NT}(t) \geq P_{min,SCR,pos,NT} \quad (45)$$

$$P_{SCR,neg,HT}(t) \geq P_{min,SCR,neg,HT} \quad (46)$$

$$P_{SCR,neg,NT}(t) \geq P_{min,SCR,neg,NT} \quad (47)$$

The result of the control reserve maximization is a scheduling concept to provide system services and enables the transmission system operator to have stable and effective control. Furthermore, if the VPP is located in lower voltage levels, this concept supports the appropriate connection system operator with defined schedule information concerning possible control reserve activation for distribution system operator. This is a necessary information during the day-ahead planning process in case of expected congestions in the distribution system.

4. Case Study

The case study demonstrates the application of the presented concepts under consideration of a test network. The impact of the operational concepts on the power system are highlighted according to the case study. The test nework in Figure 6 is based on the CIGRE Benchmark System. The used medium voltage network represents areas of central city, suburb and country grid according to [36 – 39]. Furthermore, there is only a single hand-over point to the overlay 110 kV network. Renewable generation from PV, wind and bio mass are collected in the voltage levels HV/MV, MV, MV/LV and LV in this test network.

Based on the investigation in [40] a defined amount of load (9815 kW) and installed renewable generation were randomly distributed. The main amount of RES is installed in country and suburb area and the main installed load is in the central city. Load profiles are modelled with standardized load profiles for households, business and agriculture. The installed renewable power for the VPP operation is considered to be 339.25 kW bio generation, 2126.56 kW wind generation and 7279.39 kW photovoltaic (pv, mainly in low voltage). The case study represents a scenario with high wind and photovoltaic infeed time-series for one day in 15 minute values and the operational starting point for bio power plants were expected to be 50 % of nominal power output. Figure 5 shows the uncontrolled renewable energy infeed forecast in the VPP area.

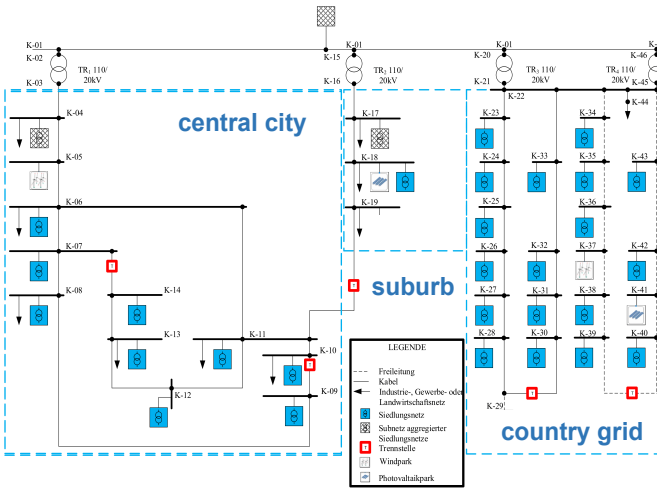


Figure 6. Medium voltage test network to [36 – 39]

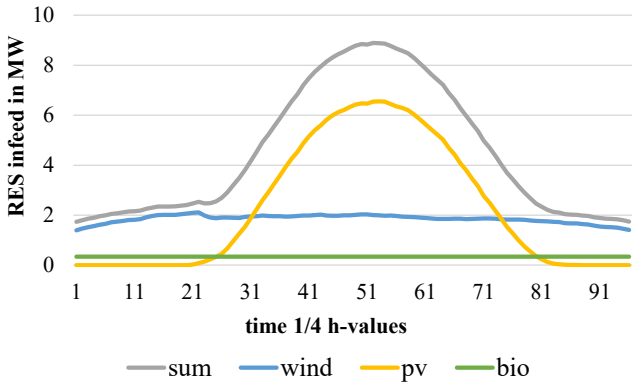


Figure 5. VPP generation day-ahead forecast

In the next step, the impact of the optimization is presented for the day-ahead planning process, with change of generation unit schedule in comparison to the uncontrolled renewable energy infeed (forecast). The test network with the implementation of load and generation as well as the according profiles offers the possibility to highlight the qualitative impact of the operational concepts on the 110 kV hand-over point exemplary. The impacts demonstrate the need of schedules and the impact of centralized controlled RES in low voltage levels. It is assumed that minimum bid and increment for control reserve is below 0.1 MW, because the case study only considers a small part of a network. The VPP itself can operate with more installed capacity of RES that is distributed in a bigger area and fulfill CR market requirements.

Furthermore, the technical oriented concepts will be exemplary represented by negative SCR contribution and negative redispatch contribution. Generation unit schedule is close to zero in case of maximum positive power provision for SCR and RD.

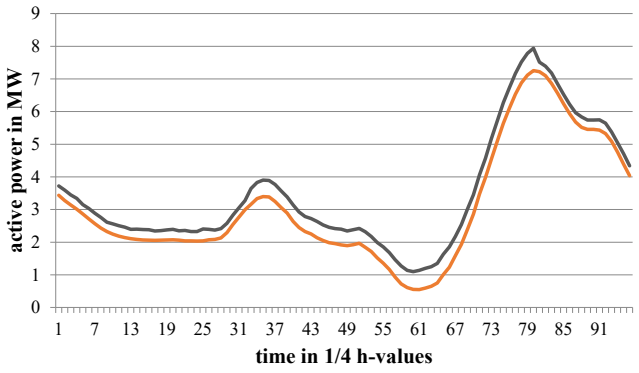


Figure 7. Active power demand of the medium voltage network with (grey) and without (orange) profit maximization

4.1. Economical oriented operation

The results of profit maximization are based on average prices for Day-Ahead market (blue) and Intra-Day market (green) as well as for CR participation (Figure 7). The revenue for PCR participation was assumed with 3646 €/MW and for SCR: 357 €/MW (pos. HT), 605 €/MW (pos. LT), 122 €/MW (neg. HT), 318 €/MW (neg. LT).

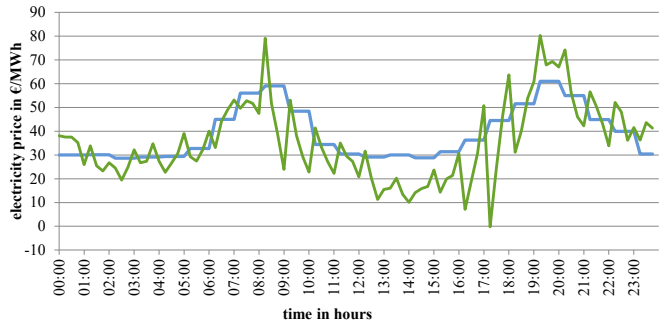


Figure 8. VPP generation unit scheduling for profit maximization

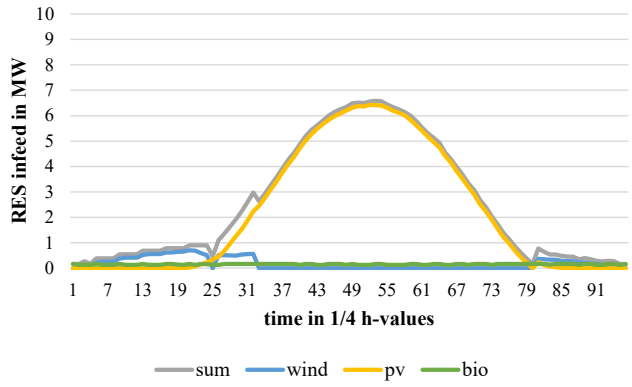


Figure 9. VPP generation unit scheduling for profit maximization

The overall expected profit for average spot-market prices of CR-market, Day-Ahead and Intra-Day market is about 1443.24 €. Most of the energy is planned to be traded on the Intra-Day market. Only less power is provided for positive SCR in low tariff with 1.563 MW and 2.297 MW in high tariff. The optimized generation unit schedule is reduced according to SCR provision (Figure 9).

Figure 7 demonstrate the impact of the profit maximization at the 110 kV hand-over point before profit maximization (orange) and after (grey). Due to the CR provision, the demand in this network area increases. It can be stated that the power demand of the network section is strongly reduced during the middle of the day, caused by a high PV infeed.

4.2. Technical oriented operation

The schedule loyalty is one of the basic needs for future system operation with high percentage of RES. The starting point for this concept is close to profit maximization, but the capacity for flexibilities has to be defined. Therefore, the day-ahead planning process is optimizing with only 95 % of forecast and a maximum of 90 % for bio generation. Therefore, the day-ahead schedule is set according to Figure 10. Furthermore, a random failure of maximum 5 % is assumed in each time step for wind and PV forecast. The Day-Ahead schedule expects a profit of about 1337.21 €, which lower compared to profit maximization. Most of the power is planned to be traded again in the Intra-Day market, but less power is provided for positive SCR in low tariff with 1.493 MW and 2.191 MW in high tariff.

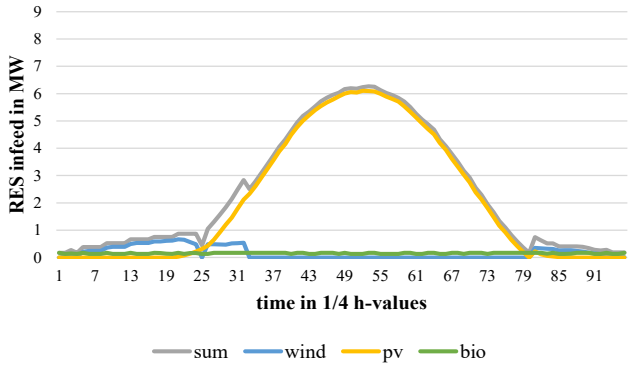


Figure 10. VPP day-ahead generation unit scheduling for schedule loyalty optimization

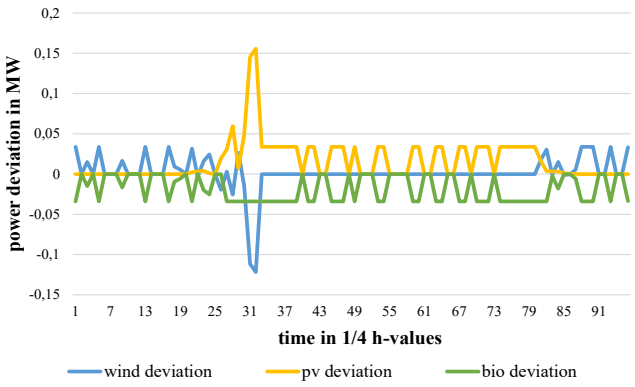


Figure 11. VPP intra-day deviation of generation in comparison to day-ahead schedule

The planned schedule can be fullfilled during the intra-day operation without any CR constraint violation and without any balancing power consumption. The intraday difference in generation unit operation is shown in Figure 11. The results demonstrate the functionality of this optimization concept. Furthermore, it was shown that the reduction of profit is relatively small in comparison to profit maximization, but the system benefit is enormous.

Eventually, the results for negative secondary control reserve and negative redispatch maximization are presented exemplary for all concepts of system service maximization. Both concepts represent the objective to maximize the power in within the planning horizon to provide maximum power that can be reduced by negative RD or negative SCR activation. Though results in generation unit schedule are the same, the reason of use is different. Figure 12 presents the day-ahead generation unit schedule. By considering the sum of generation, power output is maximized with repect to planned negative SCR and RD. In case of call of power reduction (RD) or negative SCR the maximum power is reduced. In particular at time step 25, there is a small reduction of wind caused by local weather conditions. It has a high impact on the maximization because PV infeed is low in the morning and the share of bio generation is small

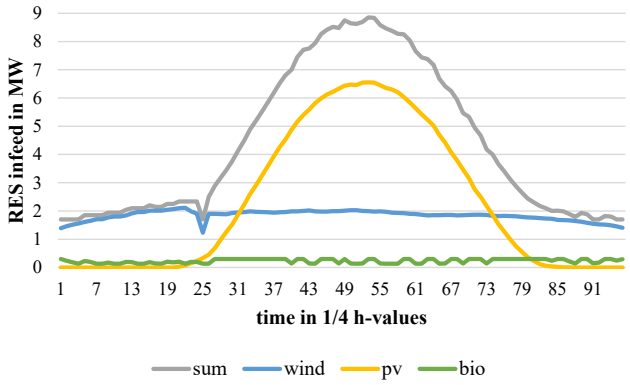


Figure 12. VPP generation unit schedule for maximization of negative SCR and negative RD

One main challenge and requirement for these concepts is the embodiment of incentives to support the system with these functionalities. On the other hand, is the concept of intra-day schedule loyalty. Schedule loyality is one of the best circumstances for system operatores point of view in terms of high voltilty during system operation. The concept of intra-day schedule loyalty needs only small incentives to be integrated in the german power system and leads only to small losses in profit. This concept can optimally be applied with controllable, forecast independent, generation or storage units. Otherwise, with a high amount of pv and wind it is necessary to make a day-ahead schedule with reduced forecast. But in the end this concept is main requirement for future system with high amount of RES and less conventionals to keep the system stable.

Figure 13 illustrate the change of power demand of the network at 110 kV hand-over point in case of call of negative SCR or negative RD (orange). The demand of active power before activation (grey) indicates a low load in comparison to the installed RES with times of power infeed in the 110 kV network. The activation of SCR or RD can almost double the demand during the middle of the day, depending on the high PV infeed.

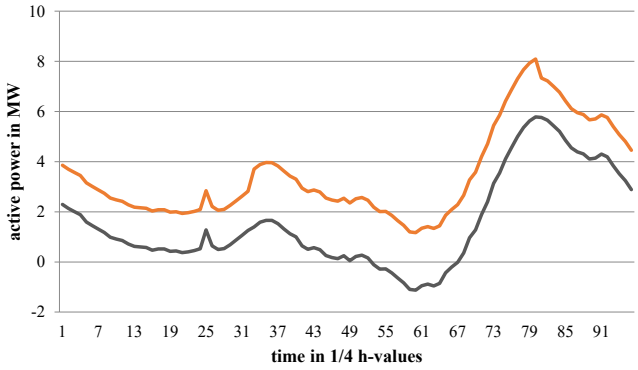


Figure 13. VPP intra-day deviation of generation in comparison to day-ahead schedule

5. Summary

The ongoing penetration of RES, especially in lower voltage levels, lead to a dramatically lack of needed system operation control funcitionalities as well as to enourmous increase of volatile generation. Therefore, this paper adresses their synergetic use in VPPs and their operational concepts. One basic advantage is the available schedule information for the system operator due to the centralized control for a defined network area. This can immensely optimize the DSO's and TSO's planning and operation process and enable generation process interventions (active and reactive power control) by involving the VPP control center, without having expensiv direct communication excess to each small unit. On top of that, it has been shown that the technical preconditions for the implementation of VPP in Germany are provided. Several suitable communication standards that meet the requirements of an intelligent power network and its decentralized generation units are already in use.Furthermore, the profit maximization concept indicates the generation of profit under currently given market conditions and without any renewable remuneration in the case study. But this result is without regarding costs for the operation of the VPP and this concept requires a sensitivity analysis to identify the minimum optimal number and kinds of generatin units to generate profit. However, there are many technical flexibilities like active and reactive power adjustments of decentralized generation units as well as storage systems, which can partly solve the uncertainty problem of demand and volatile generation forecast in particular as well as the contribution of system services. For this purpose, the concepts of intra-day schedule loyalty, control reserve maximization and redispatch maximization have been presented. The different concepts can be used in order to improve a stable and reliable system operation today and in the future. The presented technical concepts of CR and RD maximization algorithms are based on the profit maximization extended by fictious profits. This leads to the advantage that the algorithms enable the definition of needed monetary incentives to fulfill technical oriented operational concepts by optimal identification of fictious prices under changing market conditions. Furthermore, the case study demonstrates the importance of the information about location of generation units and day-ahead scheduling for CR and RD

maximization, where the call of system services can dramatically change the power demand at the hand-over point.

Eventually, the economical aspect is the most important one, especially from a VPP operator's point of view. As long as there is no sufficient level of monetary incentives, a reasonable support for the power system will not be provided. Under the given market conditions and the validity of the German Renewable Energy Act the profit for a single generation unit is much higher as for a VPP system compound, especially with regard to installation costs for a communication system and control center. Nevertheless, the future market will be dominated by renewable energy sources, so that the VPP concept will become more and more important for maintaining and improving a reliable and secure power system operation.

Limitations of the presented methods might be the computational time in case of increasing the number of constraints and by using MATLAB as the solver. For example, by defining individual technical and economical constraints for each generation unit in a VPP, the optimization complexity will increase dramatically, so that a more professional mathematical solver (e.g. Gurobi optimizer) is needed. Furthermore, the presented methods are day-ahead planning methods for a complete schedule over 96 time steps without any interruption in between two time steps. A further challenge can be to transfer the concepts of CR and RD into intra-day operation by taking system states of network frequency, line congestions and voltage stability into account [41], under consideration to calculate a new optimal solution for each next time step. By taking this aspects into account, the mathematical problem will rise strongly depending on the size of the network, size of VPP and the assumptions made.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper has been worked out as part of the project "REGEES" funded by the German Federal Ministry for Economic Affairs and Energy. The authors are grateful for funding their research.

References

[1] A. Richter; N. Moskalenko, I. Hauer, T. Schröter and M. Wolter, "Technical integration of virtual power plants into German system operation", 14th International Conference on the European Energy Market (EEM): 6-9 June 2017, Dresden, Germany, DOI: 10.1109/EEM.2017.7981876

[2] Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit, Nationale Klimapolitik, http://www.bmub.bund.de/themen/klima-energie/klimaschutz/ nationale-klimapolitik, 14.09.2015

[3] Statistisches Bundesamt, Bruttostromerzeugung 2016: 30 % stammten aus erneuerbaren Energien, 2017

[4] Informationsplattform der Deutschen Übertragungsnetzbetreiber, EEG-Anlagenstammdaten, Stand 31.12.2013

[5] Bundesministerium für Wirtschaft und Energie, Bausteine für die Energiewende: 7Eckpunkte für das 'Verordnungspaket Intelligente Netze', 2014

[6] PwC, Virtuelle Kraftwerke als wirkungsvolles Instrument für die Energiewende, PricewaterhouseCoopers AG (pwc), 2012

[7] B. Droste-Franke et al., Brennstoffzellen und Virtuelle Kraftwerke – Energie-, umwelt- und technologiepolitische Aspekte einer effizienten Hausenergieversorgung, Springer-Verlag Berlin Heidelerg, 2009

[8] Energietechnische Gessellschaft im VDE (ETG), Smart Distribution 2020 – Virtuelle Kraftwerke in Verteilungsnetzen, VDE, Frankfurt a. M., 2008

[9] Next Kraftwerke GmbH, Praxisbeispiel Virtuelles Kraftwerk: Flexibilität aus erneuerbaren Energien und Industrieprozessen, 2014

[10] Deutsche Windguard, Kostensituation der Windenergie an Land in Deutschland, Varel, Dezember 2015

[11] Fraunhofer ISE, Aktuelle Fakten zur Photovoltaik in Deutschland, Freiburg, Fassung vom 22.04.2016

[12] Fraunhofer ISE, Stromgestehungskosten Erneuerbare Energien, Freiburg, November 2013

[13] B.M. Buchholz, C. Brunner, A. Naumann and A. Styczynski, "Applying IEC standards for communication and data management as the backbone of smart distribution", Power and Energy Society General Meeting, 2012 IEEE, 10.1109/PESGM.2012.6345011

[14] CEN/CENELEC/ETSI Joint Working Group, Standards for Smart Grids, Final Report, 2011

[15] Dilan Sahin, et al., Smart Grid Technologies: Communication 1st ELECON Workshop 41 Technologies and Standards Transactions on Industrial Informatics, vol. 7, no. 4, pp. 529-540, November 2011.

[16] International Electrtechnical Commission (IEC), Electropedia: the World's Online Electrotechnical Vocabulary, http://www.electropedia.org/

[17] BDEW (Bundesverband der Energie und Wasserwirtschaft) and ZVEI (Zentralverband Elektrotechnik- und Elektroindustrie e.V.), Smart Grids in Deutschland – Handlungsfelder für Verteilnetzbetreiber auf dem Weg zu intelligenten Netzen, 2012

[18] M. Richter, "PMU-basierte Zustandsabschätzung in Smart Distribution", Dissertation, Otto-von-Guericke-Universität Magdeburg, Germany, 2016

[19] N. Moskalenko, "Optimal Dynamic Energy Management Systems in Smart homes", MaFo, vol. 59, ISBN 978-3-944722-16-0, Magdeburg, 2014

[20] A. Naumann, "Leitwarte im Smart Grid", MaFo, vol. 47, ISBN 978-3-940961-81-5, Magdeburg, 2012

[21] Wen-Yeau Chang, "A Literature Review of Wind Forecasting Methods", Journal of Power and Energy Engineering, 2013

[22] C. Skittides and W-G. Früh, "Wind forecasting using Principal Component Analysis", Journal Elsevier, Edinburgh, 2014

[23] Aoife M. Foley et al., "Current methods and advances in forecasting of wind power generation", Journal Elsevier, 2011

[24] National Renewable Energy Laboratory (NREL), Solar and Wind Forecasting, http://www.nrel.gov/electricity/transmission/resource_forecasting.html, 2015

[25] International Energy Agency (IEA), Photovoltaic and Solar Forecasting: State of the Art, Report IEA PVPS, Canada, 2013

[26] Next Kraftwerke GmbH: Strommarkt, [Online], Available: https://www.next-kraftwerke.de/wissen/strommarkt.

[27] Tariq Kamal, Syed Zulqadar Hassan, Muhammad Hussnain Riaz, Hui Li, Muhammad Sarmad and Gussan Maaz Mufti, "Design and Control of Photovoltaic/Microturbine/Super-Capacitor based Microgrid System", 20th IEEE International Multitopic Conference 2017 (INMIC 2017) at National University of Computer and Emerging Sciences (FAST-NUCES) Lahore campus, Pakistan from November 24 – 26, 2017.

[28] European Energy Exchange AG (EEX), EEX Produktbroschüre Strom, 2012. [Online]. Available: http://eex.com/blob/66450/6aee0902c1f341968d1f1b948f3b1a5b/konzept-strom-release-03a-deutsch-pdf-data.pdf.

[29] P. Konstantin, Praxisbuch Energiewirtschaft, 3 Hrsg., Berlin, Heidelberg: Springer Vieweg, 2013

[30] EPEX SPOT SE. Price List: Valid as of 1st January 2016. https://www.epexspot.com/document/34180/EPEX%20SPOT_Price%20List%20January%202016.pdf

[31] 4 German TSO, Market for control reserve in Germany, [Online] Available: https://www.regelleistung.net/ext/static/market-information

[32] 4 German TSO, About control reserve and Tender details, [Online] Available: https://www.regelleistung.net/ext/

[33] Bundesnetzagentur, Bundesnetzagentur verbessert die Bedingungen zur Teilnahme an den Regelenergiemärkten Strom, Press announcement, 28.06.2017

[34] W. Domschke, Einführung ins Operation Research. 9. Auflage, Berlin: Springer-Verlag, 2015

[35] M. Tröschel, "Aktive Einsatzplanung in holonischen virtuellen Kraftwerken", Dissertation, 2010

[36] CIGRE Task Force C6.04, Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources, ISBN: 978-285-873-270-8, 2013

[37] Netzentwicklungsplan Strom 2025, Version 2015: Zweiterer Entwurf der Übertragungsnetzbetreiber, Feb. 2016, [Online]. Available: http://www.netzentwicklungsplan.de/_NEP_file_transfer/NEP_2025_2_Entwurf_Teil1.pdf

[38] Thüringer Energie- und GreenTech-Agentur, Studie zu lokalen Energiespeicherbedarfen: Analyse und Prognose der Technologien und Anwendungsfelder thermischer und elektrischer Energiespeicher auf Nieder- und Mittelspannungsebene, Juni 2014, Erfurt
[39] BMWi, Studie: Moderne Verteilernetze für Deutschland, Abschlussbericht, September 2014
[40] A. Richter, Chr. Ziegler, N. Moskalenko and M. Wolter, Das Virtuelle Kraftwerk als gewinnorientierter Akteur mit verteilnetzunterstützendem Potential, im Tagungsband 4. Konferenz Zukünftige Stromnetze für Erneuerbare Energien, Berlin, Germany, 2017
[41] D.P. Kothari and J. S. Dhillon, Power System Optimization, Prentice-Hall of India, 2004, 2nd Edition, 2011

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 148-154 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Development of Indicators for Technical Condition Indexing of Power Transformers

Gints Poiss*, Sandra Vitolina, Janis Marks

Riga Technical University, Department of Electrical Machines and Devices, Latvia

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 08 January, 2018
Online: 30 January, 2018

Keywords :
Power transformer
Condition indexing
Fuzzy logic
Transformer windings
Oil insulation

ABSTRACT

Reliable operation of a power transformer with a certain load depends on the technical condition of individual construction parts and the ability to prevent defects that can cause a failure. During the lifecycle of a transformer, valuable data is constantly accumulated, which forms the basis for technical or risk assessment of the equipment. Therefore it also serves as a ground for the decisions on further operation, or repairs, or replacement. To achieve this goal, data need to be systematized. Since technical condition indexing allows combining various types of data including results of diagnostic tests is used within the framework of this research.

As part of a larger risk assessment methodology, algorithms for two indicators are proposed in this paper, and they are based on results of electrical measurements and analysis of oil parameters, respectively. The novelty of the algorithms for indicators introduced in this paper is based on analysis of features specific to the power system in Latvia such as large proportion of aged transformers, low loading level, significant variation in oil volumes, and statistics on typical faults. Proposed limits of parameters are verified with data from operation history. Taking into account the differences in the measurement periodicity, the indicator that is based on electrical measurements assesses the individual constructive parts of the transformer (windings and core, bushing and on-load tap changer) separately, whereas the other, indicator combines the results of oil parameters into a single assessment. These indicators were verified by using 30 transformers from the Latvian power system and the obtained results coincide well with the operation history.

1. Introduction

This paper is an extension of the results disseminated in the international conference paper relating to the development of risk indicator for power transformers based on electrical measurements [1]. This risk indicator is part of a condition indexing algorithm that allows categorizing a large volume of technical data obtained during operation and maintenance of a power transformer into discrete risk categories. It provides information in a fast and systematic way, so that transformers can be easily compared and critical ones can be selected for further investigation.

The main goal of this extended paper is to present an algorithm for determining the numerical value of the indicator based on results of electrical measurements (EM indicator), as well as for establishing the numerical value of the indicator based on results of oil parameters analysiss. The outcome for both indicators is the risk category derived on a scale from 1 (lowest risk) to 5 (highest risk, outage expected or immediate action necessary). The proposed indicators correspond to diagnostic tests regularly performed on power transformers in Latvian transmission system. However, significant differences in the periodicity of oil checks and electrical measurements determined the disparate structure of both algorithms and the mathematical methods selected.

The proposed indicators have been verified with 30 case studies. Results of diagnostic tests both from transformers that required repairing, and transformers in a normal operation condition were used. The authors express their gratitude to the Latvian Transmission System Operator (Augstsprieguma tikls, JSC) for providing data for the case studies.

Different approaches are applied to develop appropriate and efficient algorithms for technical condition index which includes such common power transformer parameters as oil characteristics, dissolved gas analysis and electrical measurements [2].

Artificial neural networks (ANN) and adaptive neuro-fuzzy inference system (ANFIS) models, for instance, are no longer limited just for diagnosing faults and aging of power transformers.

*Corresponding Author: Gints Poiss, Email: gints.poiss@edu.rtu.lv

These models can also be used to quantify the general condition of a power transformer [3], also as synthetic risk assessment methods [4] and fuzzy logic [5].

The different diagnostic tools and methods, periodicity and power system charchteristics are actually driving the need to develop a unique, expert-made assessment models for a certain power system. The novelty of the algorithms for indicators introduced in this paper is based on analysis of features specific to the power system in Latvia such as large proportion of aged transformers, low loading level, significant variation in oil volumes, and statistics on typical faults. Proposed limits of parameters are verified with data from operation history. The algorithms provide evaluation of the main constructive parts of a power transformer, and also the structure of indicator algorithms is planned as adjustable, if the amount of applied diagnostics tests is changed.

2. Technical Condition Index as a Part of Risk Matrix

Configurations of the risk matrix proposed for Latvian power system within the framework of this research is shown in Figure 1, and can be effectively used to evaluate each power transformer in the system, as well as to plan the maintenance or replacement of aged units. Risk matrix is divided into 3 parts, where green indicates a low-risk with no concern, blue indicates a moderate risk region or a transformer in a normal operation and technical condition, whereas the red region indicates a need for immediate action. The greatest attention will be paid to 4 cases depicted in Figure 1, that fall in the high-risk region.

In Figure 1, the ordinate axis of the risk matrix, shows operation characteristics and it is based on such important parameters as transformer age, load, maintenance history, the existence of monitoring system, importance in the system (as additional option). Three indicators are used to determine the technical condition index. Together with scoring and weighting factors it is depicted as a value on abscise axis of the risk matrix. Parameters used to determine the technical condition index are given in Table 1 and they correspond to diagnostic tests regularly performed on power transformers in Latvian transmission system.

The algorithm of the indicator based on dissolved gas analysis (DGA) is provided in [6]. It is based on assessment of 7 key gases shown in Table 1 and includes analysis of features specific to the power system in Latvia where power transformers are aged and variations in oil volumes are significant. Transformer loading, operation of the on-load tap changer (OLTC), and oil treatment are also taken into account in order to quantify this indicator.

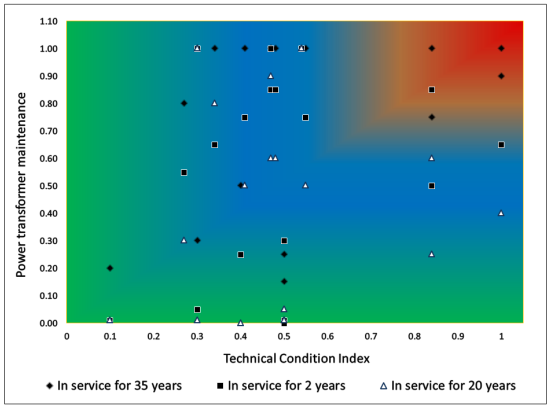


Figure 1. Proposed configuration of risk matrix

The indicator based on electrical measurements for technical condition indexing of a power transformer is developed by evaluating separately 3 main parts of a power transformer: windings, bushings and OLTC. A failure of any of these parts may be critical for a power transformer, the surrounding equipment, environment and even service personnel. The indicator based on analysis of oil parameters such as flash point, dissipation factor, acidity and moisture content is developed as a combined approach of binary and fuzzy logic which allows evaluating the risk on basis of several oil parameters as a single entity.

Both indicators are described in detail in the following chapters of this article.

Table 1. Indicators for technical condition indexing

Electrical measurements		
Windings and core	***Bushings***	***OLTC***
Power factor, PF	Power factor, PF	Static resistance, SR
Insulation resistance, R_m	Insulation resistance, R_m	Transition time & current ripple from dynamic resistance measurement (DRM)
Short-circuit impedance, Z_k	Capacitance, C1	
No-load losses, P_0		

Transformer oil	
Oil analysis	***DGA***
Dissipation factor	H_2, CH_2
Flash point	C_2H_4 C_2H_6
Moisture content	C_2H_2
Acidity	CO, CO_2

3. Development of Electrical Measurement Indicator

3.1. Winding and core

Power factor, insulation resistance and short-circuit impedance allow verifying the technical condition of transformer windings, whereas variations in no-load losses reflect the condition of a transformer core.

A flowchart of a multi-step algorithm is shown in Figure 2. where scoring system for windings and core as a part of the EM indicator is derived on a scale from T=1 (lowest risk) to T=5 (highest risk, outage expected).

The first step is input of measured parameter values from test reports: insulation resistance (R_m), power factor (PF), short-circuit impedance (Z_k), no-load losses (P_0), and winding temperature readings from a thermometer on a transformer.

Next step is data processing. As a reference point, it requires data from the factory test report for a particular power transformer, as well as background information such as age, rated voltage, etc. This information is stored in a specific technical file that is prepared for each transformer in the system. This file is also used for other condition indexing indicators. Since the technical file contains limits for the evaluation of parameters shown in Table 2, modifications are restricted.

During data processing the input data are modified into a suitable form for further analysis. PF values from the test report are calculated at base temperature of 20℃. Thus the limits proposed in Table 2 and stored in the technical file can be applied for risk evaluation (E1 indicates low risk, and E5 – high risk). PF limits are based on recommendations provided in the standard IEEE C57.152-2013. On basis of maintenance experience, additionally value for level E5 were specified. PF=1.5% at 20℃ is an effective parameter for detecting high risk since moisture and contamination of windings in this case in practice proved to be too high for reliable operation. Another parameter used for high risk detection is the D_t coefficient. It is a ratio between the measured resistance and rated voltage of high voltage winding (R_m/U_n). Coefficient D_t=1MΩ/kV at base temperature of 20℃ is set as the minimum insulation resistance limit [7]. For further analysis, values from the factory test report R_{fact} adjusted to base temperature are used as reference to evaluate if the decrease of measured R_m is acceptable.

Limits for evaluation of short-circuit impedance Z_k are based on [8] and given in Table 2. Two values of Z_k between phases should be within E4 of each other. Lager deviation can indicate on mechanical deformation of windings. A single-phase connection no-load losses P_0 measurement is a test that has been historically used for evaluating the magnetic system of power transformers in Latvia. A specific feature is that different test voltage is used (either 220V or 380V). It causes difficulties if evaluation is based on traditional comparative analysis to a factory test. In order to use results of a single measurement in this algorithm, limits based on statistical data and recommendations of [9] and performed study of measurement history of no-load losses from 100 transformers [1].

3.2. Bushings

Electrical measurement results can provide information about degradation of bushing insulation and help to avoid potential failures, such as bushing explosion and transformer winding deformation. Flowchart for bushing assessment is given in Figure 3, and evaluation process is similar to the windings assessment described before. Although a 3-level scoring system is proposed as more suitable for bushings as more polarizing. It starts with evaluation of results of PF measurements for bushings by applying respective limits from Table 2. for data from test reports are used for input. The input is followed by the next step – data processing where all data are modified in a suitable form just like in the algorithm for windings and the core part.

Table 2. Limits of parameters included in EM indicator

Parameter	Level			
	E1	E3	E4	E5
Windings and core				
PF, %, at 20℃	<0.5	0.5	1	≥1.5
$P_{0\ phase\text{-}phase}$, %		$P_{0\ A\text{-}B}$>40 $P_{0\ A\text{-}B}$<20 $P_{0\ A\text{-}C}$>4		
$Z_{k\ phase\text{-}phase}$, %			≥3	
Bushings				
PF, %, at 20℃	<0.5			≥1
On Load Tap Changer				
$SR_{ph\text{-}ph}$, %	<2		>2 (at least in 3 taps)	≥5
Transition time, t_t, ms	<100			≥200
$I_{N1,2}$, %			60	

If the measured PF value is above the limit E5 as it shown in Table 2 or the rated capacitance C1 of a bushing differs from factory measurement by more than 10 %, to reduce the risk of failure, it is advisable to remove the bushing from service [10].

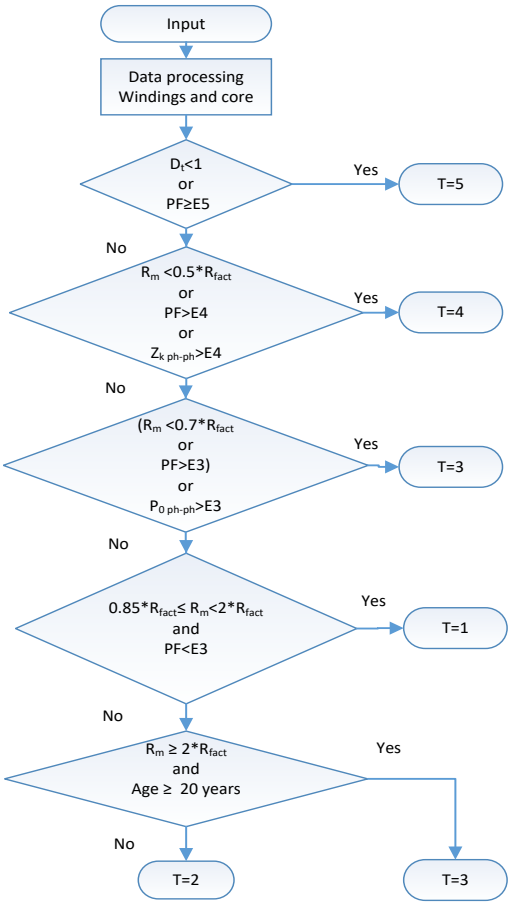


Figure 2. The algorithm of EM indicator for windings and core

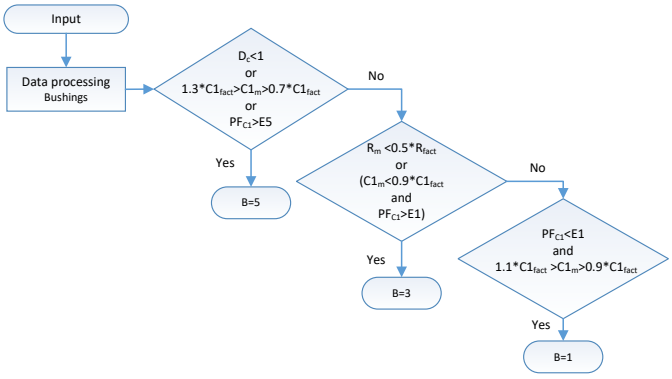


Figure 3. The algorithm of EM indicator for bushing

3.3. OLTC

The static resistance measurement (SRM) is an important tool in this industry and it is used to check for loose connections, broken strands, as well as high contact resistance in tap changer. For decades, SRM is performed for each tap in the tap changer for all power transformers in Latvia and the experience is significant. Therefore SRM is included in the OLTC algorithm as the exclusion rule which can only indicate if the risk is low or high.

For input in Figure 4, SRMs of all phases and all taps are assessed without the factory test report results; therefore the technical file is not needed.

In accordance with [11], the agreement with limit E5 indicates on high risk and it is an important signal to consider the need to repair tap changer immediately. If the SRM difference between phases is within the limit E1, it indicates on low risk. If resistance in 3 taps exceeds the limit E1, it should be investigated and TC=4 is assigned.

Dynamic resistance measurement (DRM) allows detecting defects caused by contact coking, contact wear, oil-film layer deposition, deviating transition times, contact timing problems, maintenance errors and damaged transition resistors. Therefore as additional criterion for the OLTC assessment – the dynamic resistance measurement (DRM) is proposed [1] evaluating two parameters form test report: transition time, ms, and current ripple change $I_{N1,2,}$ %. The proposed limit values E1 and E5 shown in Table 2 are assigned on basis of [12] research and were verified by authors with analysis of 76 test reports of DRM.

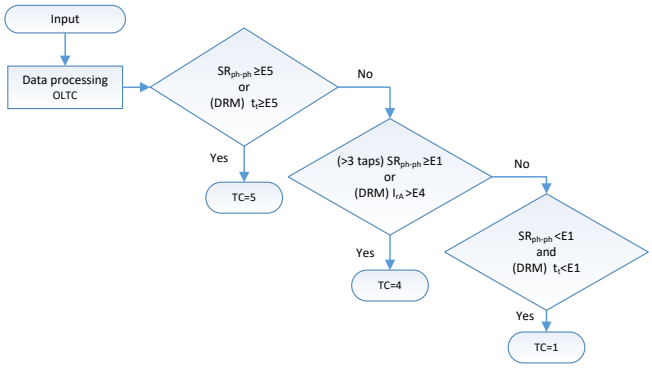


Figure 4. The algorithm of EM indicator for OLTC

4. Development of Oil Indicator

4.1. General description of oil indicator algorithm

Generally mineral transformer oil is used in power transformers in power system of Latvia. Regularly measured oil parameter are dielectric strength, moisture content, acidity, dielectric dissipation factor, interfacial tension, viscosity, flash point, and sludge content. Method for assessment of the measurement results described in the national standard involves comparison of data obtained in laboratory with the given assessments limits, and conclusions are obtained for each oil parameter separately rather than as a single entity. The standard does not provide guidance on how to obtain one conclusion based on multiple contradicting input parameters.

Fuzzy logic is widely applied for dissolved gas analysis (DGA) of power transformers [13, 14] since DGA contains relatively many input parameters. If binary logic is used a contradiction can occur, where some of the input parameters correspond with a good condition but others with a bad condition. Fuzzy logic can provide more detailed assessment of technical condition. For example, fuzzy logic based algorithm with acidity and interfacial tension of oil as input parameters is used to estimate remnant life of a power transformer [15]. Similar approach of data processing is described in [16] in which remaining service life of transformer in years is predicted based values of viscosity, resistivity, particle count, acidity and moisture content of oil. Another research [17] shows the possibility to combine measured values of moisture and acidity of oil, and the power factor of the winding to determine the type of transformer defect based on the physical and chemical condition of insulation in cases with contradicting input data values.

Therefore within the framework of this research fuzzy logic as data processing method is used to develop a technical condition indicator for a power transformer based on measurement results of oil parameters. Four independent oil parameters (flash point, dissipation factor, acidity and moisture content) and additional parameter (changes in flash point value in two subsequent measurements) are used as input data as shown in Figure 5. Limits shown in Table 3 are based on typical values observed for transformers installed in transmission network in Latvia. As a result, algorithm calculates the value of transformer oil indicator within limits from K=1 – (low risk) to K=5 (high risk, immediate actions required).

Flash point (FP) is the lowest temperature at which oil in certain conditions releases such amount of vapour that, together with air, creates a flammable mixture leading to fire hazard. Therefore, this parameter is proposed for indicating high risk. Based on practical experience in decision-making for transformer repairs dissipation factor (tanδ) is proposed as another parameter indicating high risk since it is sensitive to oil aging, moisture, as well as changes in contamination levels.

Table 3. Proposed Limits for Oil Parameters

Flash point, °C	Upper limit	Acidity, mgKOH/g	Moisture content, ppm	Tanδ, %
125 or ≥Δ5 in 2 subseqent measurements	Level 1	<0.05	<5	<1
	Level 2	≤0.075	≤10	≤2.5
	Level 3	>0.1	>15	>4

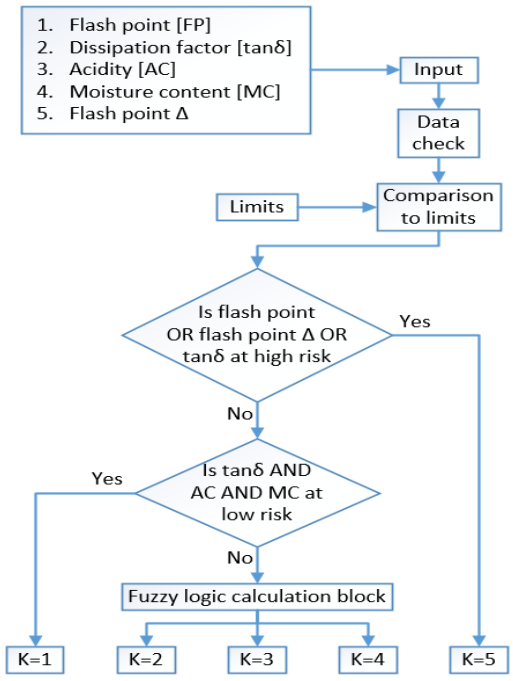


Figure 5. Overall flowchart of oil indicator algorithm

Dissipation factor tanδ along with acidity and moisture content are proposed as criteria for medium risk evaluation. In order to reach a unified conclusion regarding the correlation between these independent oil parameters, the use of fuzzy logic is suggested.

The operation of the algorithm begins with reading the input parameters, namely results from the most recent oil measurement which then are stored in technical file that is prepared for each transformer in the system for possible repetitive usage. If any of the necessary oil parameters have not been measured in this particular testing, the algorithm assumes the value of that parameter as in Level 2 (see in Table 3).

If the value of the parameter is repeatedly unknown, the approximation is performed in a logarithmic scale as shown in Figure 6 where approximated value aims for the Level 3, yet cannot reach it, since the difference from the maximum boundary is halved per iteration.

Figure 6. Assumed values for unknown parameter value

Next step of the proposed algorithm is the evaluation of exclusion cases indicating high risk (either flash point or changes in value of flashpoint or dissipation factor exceeds the limits of Level 3) or low risk (values of dissipation factor and moisture and acidity are all below limits of Level 1). Fuzzy logic calculation block activates if neither of the exception cases checks included in the algorithm is positive.

4.2. Fuzzification

Developed shape and slope for membership functions used within fuzzy logic calculation block of proposed algorithm is shown in Figure 7 originally, the selection of membership function shape and slope was carried out for each of the three oil parameters. Illustration of membership function for tanδ is shown in Figure 8. The triangle and trapeze forms were tested for M (medium risk) membership function, whereas slope was altered for L (low risk) and H (high risk) membership functions. All variations were tested by using the results of oil tests for 10 transformers with given maintenance history and different technical condition.

Results show that output values are not affected by the shape of M membership function, however, in some cases value changes were observed by alteration of the slope of L and H membership functions. After evaluation it was concluded that the combination of increased H membership function effect has the best correlation with technical condition based on maintenance history.

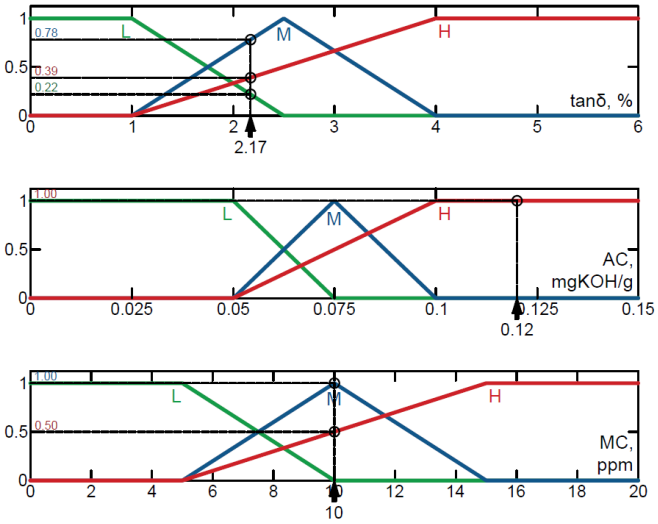


Figure 7. Membership functions of oil indicator

For obtaining corresponding output values, a rulebox is created as shown in Table 4. It contains all possible combinations. Altogether, there are 27 combinations, since for each of the 3 input parameters used in the fuzzy logic block 3 membership functions are assigned. The specific weight value is the minimum of all three membership values for a given combination. The specific weights are summed together for each result value and thus a fuzzified output is obtained. Afterwards, the gravitation centre of this output function is calculated. The obtained result is more precise, as more points are used for the calculation.

4.3. Fuzzy inference and defuzzification

An example is illustrated in Figure 7 with following values of oil parameters as input data from regular oil sample test of a particular 110 kV transformer: flash point, 139°C; changes in flash point value between two subsequent measurements, 2°C; dissipation factor tanδ, 2.17%; acidity, 0.12mgKOH/g; moisture content, 10 ppm. Since exception cases included in the algorithm are negative in this case fuzzy logic calculation block activates.

Table 4. Rulebox for fuzzy logic calculation block

#	Input Variables			Corresponding output	#	Input Variables			Corresponding output	#	Input Variables			Corresponding output
	tan δ	*AC*	*MC*			*tanδ*	*AC*	*MC*			*tanδ*	*AC*	*MC*	
1	L	L	L	2	10	H	M	L	3	19	L	M	H	3
2	M	L	L	2	11	H	L	M	3	20	M	M	H	3
3	L	M	L	2	12	H	M	M	3	21	H	H	L	4
4	L	L	M	2	13	L	H	L	3	22	H	H	M	4
5	M	M	L	2	14	M	H	L	3	23	H	L	H	4
6	M	L	M	2	15	L	H	M	3	24	H	M	H	4
7	L	M	M	2	16	M	H	M	3	25	L	H	H	4
8	M	M	M	2	17	L	L	H	3	26	M	H	H	4
9	H	L	L	3	18	M	L	H	3	27	H	H	H	4

The membership of dissipation factor is 0.78, 1 for acidity and 0.5 for moisture content. The specific weight is calculated as 0.5. This process is repeated for each matching combination. The gravitation centre of this output function is calculated, as well as the position of gravity centre on x axis is estimated as shown in Figure 9. The third of the position of gravity centre is the result and final output, in this case K=4 for oil indicator is obtained, which indicates a rather high operation risk.

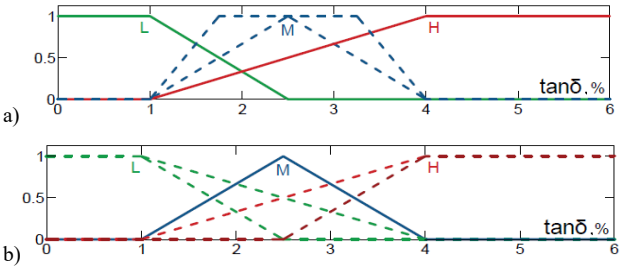


Figure 8. Selection of the shape and slope for membership functions for tanδ: a) shape for M membership function; b) slope for L and H membership function

	Membership function	Membership value
tanδ	L	0.22
	M	0.78
	H	0.39
AC	L	0
	M	0
	H	1
MC	L	0
	M	1
	H	0.5

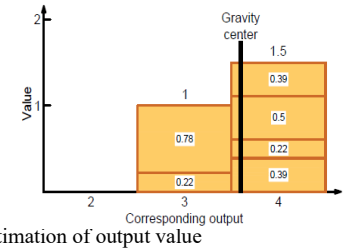


Figure 9. Estimation of output value

5. Verification of EM and Oil Indicator

Results of measurements within a time period of 8 years for 30 power transformers are used to verify both indicators and see how they complement each other. It has to be noted that the power transformer park in Latvia is aged and power transformers installed between 1967 and 2000 were used for verification with different technical condition, repairing plans and failure rate. Figure 10 shows numerical values of EM and oil indicator obtained for the latest available measurement set of each transformer. Since polarized score is used omitting the values 2 and 4 for bushings and 2 and 3 for OLTC the combined score 3&4 is allocated in a separate column. Mainly a numerical value of 1 was assigned to parameters OLTC and bushings which reflects timely scheduled repairs. In those two cases where score value of 5 is assigned for bushings immediate decisions to replace them have been made by system operator. Similarly OLTC is repaired immediately if there1 is a variation from normal operation or a problem is detected.

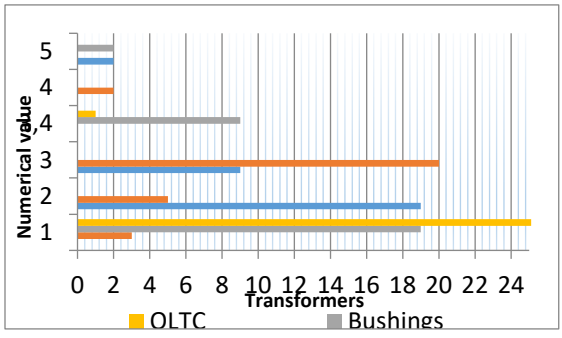


Figure 10. Numerical values of indicator parameters of 30 power transformers

Values 2 and 3 are obtained most frequently for oil analysis most showed result 2 and 3, which indicates either normal operation or marks out necessity for small maintenance works such as the change of silicagel in a thermosyphon. In two cases when score of 4 for winding and core is assigned result dramatic decrease of winding isolation resistance and increase of power factor therefore obtained results concise well with operation history.

For further analysis 4 cases from 30 were selected to illustrate how indicators together reveal the technical condition of a power transformer. Cases 1 to 3, plotted in Figure 11, depicts the latest measurements available for three different transformers. Case 2 reflects a transformer in an almost perfect technical condition. But in case 1, problems with OLTC can be noticed and analysis of measurement history revealed increased static resistance in 3 taps. Case 3 reveals bushing defect, caused by decreased insulation resistance in scheme C2 and increased power factor in scheme C1.

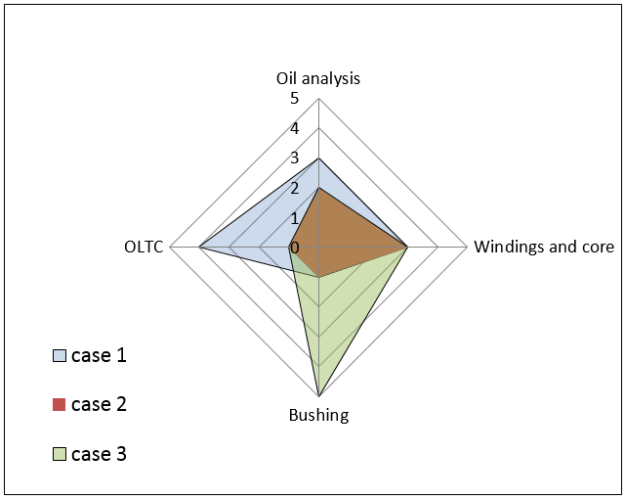


Figure 11. Case studies results

Since electrical measurements are performed on average once in four years EM indicator is more static. Case 4, plotted in Figure 12, illustrates the role of the oil indicator as the first implication of a possible fault in a transformer since its results are more dynamic. It can be observed that technical condition of this particular transformer has decreased due to increase in moisture level over the years.

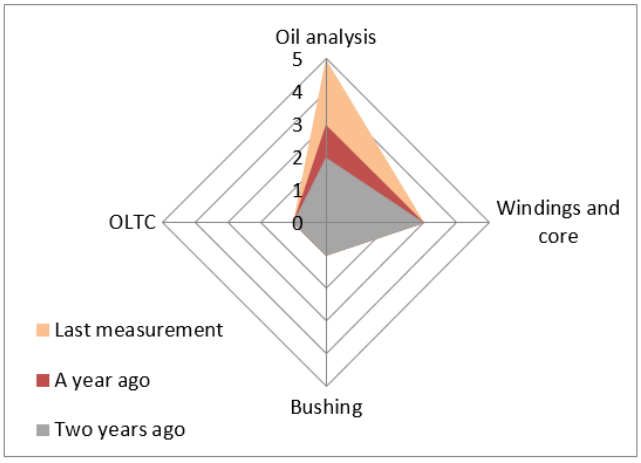


Figure 12. Oil analysis results

6. Conclusions

Literature review leads to a conclusion that methods for risk assessment of power transformers are mainly based on the results of performed diagnostic measurements and operation history. Since maintenance strategy may vary over the time and may depend on decisions by a particular transmission system operator the development of a modified risk matrix is a common global practice.

Configuration of risk matrix for power transformers in Latvian transmission network proposed within the framework of this research is composed of three regions indicating low risk, moderate risk and high risk the last one requiring immediate action. The place of a transformer within the risk matrix depends on its technical condition index (x-axis) and such parameters of operational history as transformer age, load, maintenance history, importance in the system (y-axis). Technical condition index is established by calculating the numerical values of the three indicators and applying a scoring and weighting factor.

To determine numerical values of the indicator based on oil analysis combined approach of binary and fuzzy logic proved to be successful since it allows evaluating risk based on several oil parameters as a single entity. Results from a verification confirm that higher output values are shown by transformers that have already attracted attention in practice. As oil analysis is done quite often, changes in values of this indicator are first indication of a possible defect in a transformer and they substantiate the necessity for further testing. However, an indicator based on results of electrical measurements more effectively serves to detect faulty constructive parts more effectively and can be used as a basis for decisions on replacement of a bushing or repairs of an OLTC.

Acknowledgment

This paper has been partly supported by the State Research Program „LATENERGI".

Conflict of Interest

The authors declare no conflict of interest.

References

[1] G. Poiss, S. Vitolina, "Development and Implementation of Risk Indicator for Power Transformers Based on Electrical Measurements", 18th International Scientific Conference on Electric Power Engineering (EPE), Czech Republic, Kouty nad Desnou, 17-19 May, 2017, pp.425-428.

[2] J. Haema, R. Phadungthin, "Development of Condition Evaluation for Power Transformer Maintenance", 4th International Conference on Power Engineering, Energy and Electrical Drives, Istanbul, Turkey, 13-17 May 2013.

[3] Hamed Zeinoddini-Meymand, Behrooz Vahidi, "Health index calculation for power transformers using technical and economical parameters ", IET Science, Measurement & Technology, Volume: 10, Issue: 7, 10 2016.

[4] Dun Lin, Yao-Yu Xu, Yu Liang, Yuan Li, Ning Liu, Guan-Jun Zhang, "A Risk Assessment Method of Transformer Considering the Economy and Reliability of Power Network", 1st International Conference on Electrical Materials and Power Equipment (ICEMPE), Xi'an, China, 2017.

[5] Juan. P. Lata , Diego. P. Chacón-Troya , R. D. Medina, "Improved tool for power transformer health index analysis", IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), 2017.

[6] G.Poiss, "Development of DGA Indicator for Estimating Risk Level of Power Transformers," 17th International Scientific Conference on Electric Power Engineering (EPE 2016), Prague, 16-18 May, 2016.

[7] A Stitch in Time: The Complete Guide to Electrical Insulation Testing, Dallas: Megger, 2006.

[8] Scope and norms of testing electric equipment (in Russian), RAO "UES of Russia", RD 34.45-51.300-97, 1997.

[9] Kaganovich E. A., Reichman I. M., Power transformer tests up to 6300 kVA and voltages up to 35 kV, (in Russian), Energia, 1980.

[10] Electric Power Transformer Engineering, Third Edition, James H. Harlow CRC Press, May 16, 2012 by CRC Press 693 pages.

[11] "IEEE Guide for Diagnostic Field Testing of Fluid-Filled Power Transformers Regulators and Reactors," IEEE Std. C57.104, 2013.

[12] J. J. Erbrink, E. Gulski, J. J. Smit, R. Leich, B. Quak, and R. A. Malewski, "On-load tap changer diagnosis-an off-line method for detecting degradation and defects: Part 2," IEEE Electr. Insul. Mag., vol. 27, no. 6, pp. 27-36, 2011.

[13] A. Abu-Siada, S. Hmood, and S. Islam, „A New Fuzzy Logic Approach for Consistent Interpretation of Dissolved Gas-in-Oil Analysis," IEEE Transactions on Dielectrics and Electrical Insulation, vol. 20, pp. 2343 - 2349, 2013.

[14] B. Nemeth, S. Laboncz, and I. Kiss, „Condition Monitoring of Power Transformers using DGA and Fuzzy Logic" on Proc. 2009 IEEE Electrical Insulation Conference, Montreal, Canada, May 31-June 3, 2009, pp. 373-376

[15] S Forouhari, and A Abu-Siada, „Remnant Life Estimation of Power Transformer Based on IFT and Acidity Number of Transformer Oil" on Proc IEEE 11th International Conference on the Properties and Applications of Dielectric Materials (ICPADM), Sydney, Australia, July 19-22, 2015, pp. 552-555.

[16] A. K. Kori, A. K. „Sharma, and A. K. Singh Bhadoriya, Intelligent Diagnostic Method for Ageing Analysis of Transformer," Energy and Power Engineering, Vol. 4 No. 2, 2012, pp. 53-58

[17] W. C. Flores, E. E. Mombello, J. A. Jardini, G. Ratta, and A. M. Corvo, „Expert system for the assessment of power transformer insulation condition based on type-2 fuzzy logic systems," Expert Systems with Applications, vol. 38, pp. 8119-8127, 2011.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 155-159 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Influence of supply frequency on dissipation factor measurement and stator insulation diagnosis

Cyrille Caironi*, Bernhard Fruth[1], Detlef Hummes[2], Rudolf Blank[3]

[1]*PDSS, Zurich, SWIZERLAND*

[2]*American University of Kuwait, Safat, KUWAIT*

[3]*B2Electronic, Klaus, AAUSTRIA,*

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 08 January, 2018
Online: 30 January, 2018

Keywords:
Dielectric
Tangent Delta
Frequency measurement

ABSTRACT

This paper is an extension of work originally presented in EIC-2017. It deals with influence of the supply frequency for dissipation factor measurements, mainly for tests under power frequency and low frequency. After a theoretical reminder, we present some experiments on single coils and stators of high voltage motors. Finally, we discuss the results and the desirability of choosing one or the other of these methods.

1. Introduction

Dissipation factor, or tan δ, is currently used for the insulation diagnostics. While the tan δ measurements are mainly performed at power frequency, some methods based on the frequency steps [1] and very low frequency have been developed. The methods using very low frequency have some advantages in term of portability, power requirement, etc. but it is interesting to see the impact of the frequency on the quality of measurement and, subsequently, of the diagnostic.

This paper presents the influence of the supply frequency on tangent delta measurement. The paragraph II reminds the theoretical bases. The paragraph III presents some experiments, firstly on single coil and secondly on a stator of high voltage motor. Test are performed at 0.1 Hz and 50 Hz. The analysis of the result presents some difference, especially on the sensitivity at relaxation phenomena. These differences are discussed, and we can define some interests on insulation diagnostic.

2. Theory reminder

Dielectric losses are currently measured and used for electrical machines diagnostics. To perform the dielectric loss measurement, the stator is considered as a capacitor and the theory supporting these measurements is based on the complex permittivity. Of course, this capacitor is imperfect.

It is schematized by the capacitance with a resistance in parallel, both in series with a resistance and an inductance, in our case, we can neglect Ls (fig. 1).

The real part of impedance is due to the series and parallel resistances Rs and Rp. Rs becomes important at high frequency when Rp has a dominating effect essentially at low frequency.

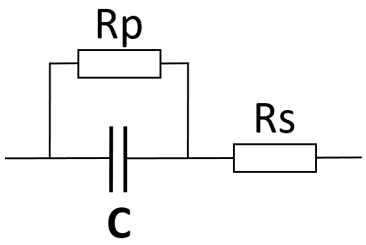


Fig. 1: Real capacitor model *(without Ls)*

*Cyrille CAIRONI, CAIRCY, 22, rue de Ceintrey 54160 Pulligny France Email : cyrille.caironi@caircy.org

Rp is the picture of the losses due to the dielectric polarization, the dielectric resistance, leakage current and also charge redistribution phenomena. All these phenomena are involved at low frequency especially at industrial frequencies and lower. De facto we currently use the model within resistance in parallel which excludes Rs (1).

$$\underline{I} = j\omega\underline{\varepsilon}C\underline{U} = \underline{I_{Rp}} + j\underline{I_C} \quad (1)$$

Due to the phenomena of relaxation and losses, the insulating materials permittivity is complex (2)

$$\underline{\varepsilon} = \varepsilon_0 \left(\varepsilon_r' - j\,\varepsilon_r''\right) \quad (2)$$

In this arrangement, ε_r' is the component which defines the capacitive current (3) and $\varepsilon_r''$is the component which defines the current I_{Rp} corresponding at the active losses in phase with the voltage (4).

$$\varepsilon_r' = \frac{\underline{I_C}}{j\omega C\underline{U}} \quad (3)$$

$$\varepsilon_r'' = \frac{\underline{I_{Rp}}}{\omega C\underline{U}} \quad (4)$$

The tangent delta, or dissipation factor is defined as (5):

$$tan\delta = \frac{\underline{I_{Rp}}}{\underline{I_C}} = \frac{\varepsilon_r''}{\varepsilon_r'} = \frac{1}{\omega CRp} \quad (5)$$

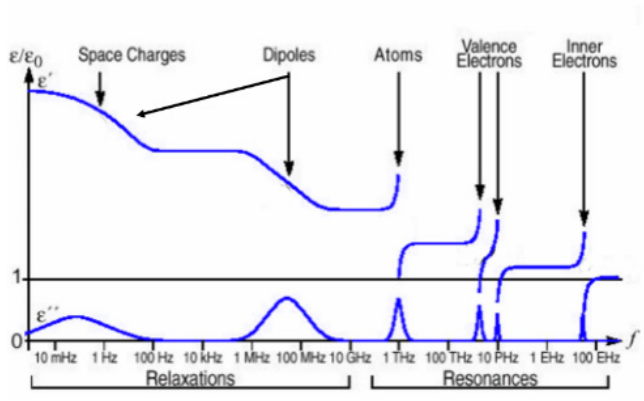


Fig. 2: effect of resonance and relaxation on permittivity *(source [2])*

Of course, we can see that the tangent delta is inversely proportional to the frequency. But the main interest of the low frequency is the physics of the material. The polarization and relaxation phenomena involved in dissipation of the insulation have different origins and each of them approximately can be linked to frequency (fig. 2) as presented in [2].

The electronic and ionic polarization have a high frequency level (around 10^{15} Hz for electronic polarization) this type of polarization is not interesting for the machine diagnostic in first approach. On the other hand, the relaxation phenomena linked to dipoles, space charges and interface charges are very interesting.

These relaxation phenomena enable us to check precisely details like pollution and polymerization, which are visible at lower frequencies (less than 1 Hz) by tan δ analysis.

Furthermore, the temperature influence is a particularly important factor as presented on figure 3 and figure 4 [3] and explained in [4].

It could be an advantage but also a risk because a variation on the temperature has a bigger impact at low frequency (less than 1 Hz) than at industrial frequency.

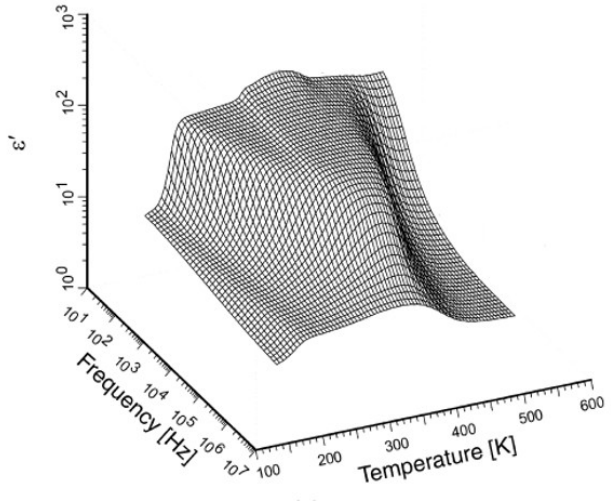


Fig. 3: Typical real part of the permittivity of porous glass versus frequency and temperature *(source [3])*

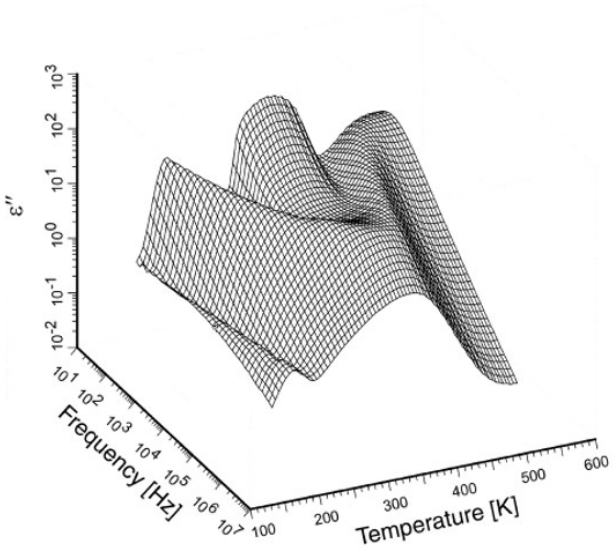


Fig. 4: Typical imaginary part of the permittivity of porous glass versus frequency and temperature *(source [3])*

3. Experimental studies

IEEE 433 recommends performing very low frequency tests at 0.1 Hz [5] and the monitoring present on the market uses this frequency. Therefore, the tests are performed at frequency of

0.1 Hz and 50 Hz, in accordance to IEEE 433 [5] and to IEC 60034-27-3 [6].

3.1. Experiments on coils

Tan δ measurements were performed on sample coils (7 kV rated) in steps of 1 kV from 1 kV to 9 kV. Measurements were performed successively at 0.1 Hz and 50 Hz.

These tests show higher tan δ values with a 0.1 Hz frequency supply than with a 50 Hz supply (fig. 5). Therefore, these measurements confirm some explanations of the theoretical reminder and they are also in accordance to the related literature and different diagnostic methods proposed in [1].

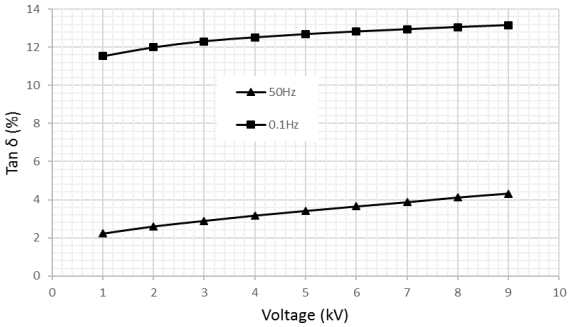


Fig. 5: tan δ versus voltage on single Coil

The difference of levels can be explained by a dielectric relaxation, which means a lowering of the dielectric losses with frequency, therefore 0.1 Hz is more sensitive than 50 Hz.

The underlying physical mechanism is usually related to the polymerization degree of the impregnating resin and interfacial polarization [4].

The voltage dependence of the losses of the sample is very similar for both frequencies. At this step, we could consider that the frequency just influences the level of tan δ. In fact, the influence is also on the behavior of the tan δ versus voltage.

3.2. Experiments on Stator

Measurements were performed on a 1.5 MW, 11 kV motor in accordance to IEC 60034-27-3 [6] and customer requirements (fig. 06).

Capacitance measurements at 50 Hz and 0.1 Hz have a similar behavior (fig. 08), on the other hand, the tan δ has a different shape at 50 Hz and 0.1 Hz (fig. 07)

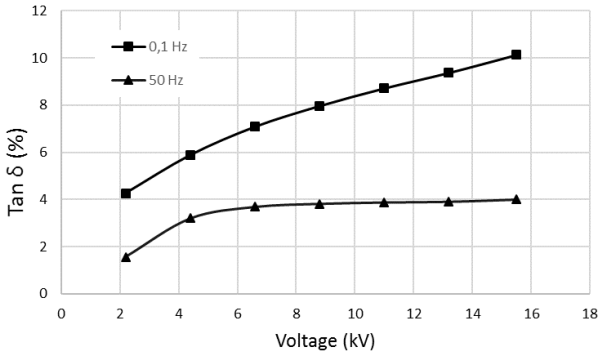
As for a simple coil, the tan δ is higher at 0.1 Hz than at 50 Hz but of different shape (fig. 07). The measurement at 0.1 Hz shows a voltage dependence which is not present at 50 Hz.

This nonlinear voltage dependences of rotating machine insulation properties typically stem from partial discharges and currents in endwinding corona protection.

Evidence of the structural influence on the dielectric properties is shown in [4] as mechanical losses and dielectric losses largely coincide. This indicates that the 0.1 HZ test method can provide information about insulation quality, e.g. brittleness, shrinkage and risk of resin cracking caused by post-polymerization

Partial discharges analysis, and especially PRPD, can complete the tan δ information. Of course, the tan δ and PRPD must be performed at the same frequency.

Fig. 6: VLF supply and measurement instrumentation on 11 kV Stator



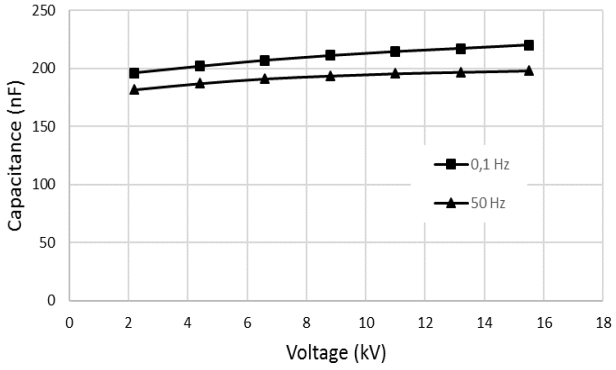
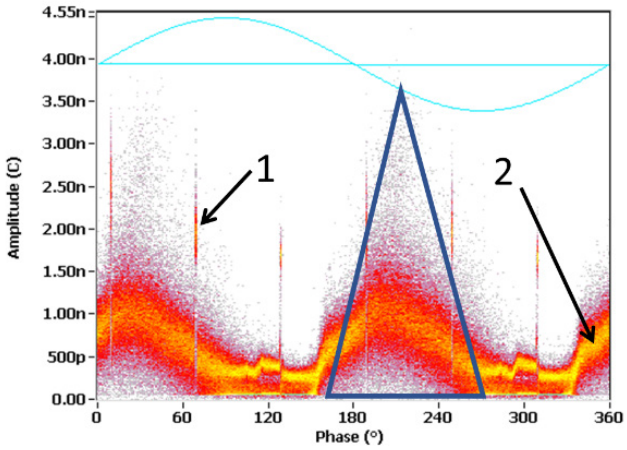


Fig. 8: capacitance versus voltage on 11 kV Stator

The patterns show solely the ionization phenomena and must be analyzed in accordance with the tan δ measurements. Furthermore, the noise phenomena have shown more important

using power frequency supplies. Phenomena linked to the noise can be solved by numerical tools presented in [7] and [8] especially in case of on-line PD monitoring.

Fig. 09 shows PRPD pattern at the network power supply and Fig. 10 shows PRPD pattern at 0.1 Hz. Despite differences, both patterns give similar information, but the 0.1 Hz can be easily compared to the tan δ.



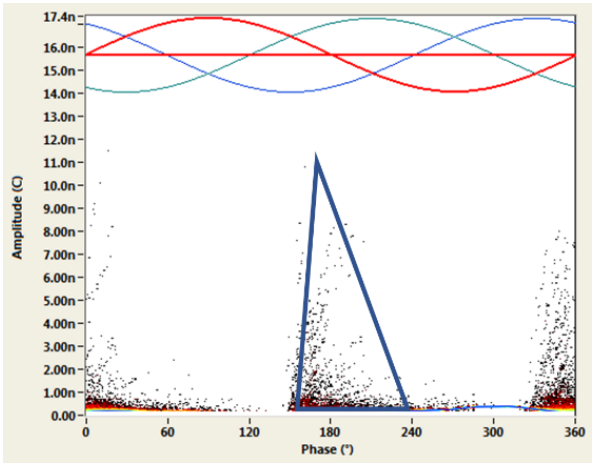


Fig. 10: PRPD at 0.1 Hz on 11 kV Stator

4. Extension to the Stator Insulation Diagnostic

The paragraph II confirms the theoretical reminder presented in paragraph I. Especially on complete stators, tan δ measurement allow to see some more phenomena at 0.1 Hz, being undetectable at 50 Hz.

In fact, the measurements at 0.1 Hz should be most efficient to the diagnostic phenomena linked to the complete insulation system [9]. Furthermore, the low frequency usage is not new. For example, measurements for CIGRE [10] was performed at 1Hz in 1990 in order to:

- do not have distortion of sinus wave (which is not perfect when it is provided by transformer at supply frequency)
- reduce the pulse pile-up
- have a cleaner PD pattern
- the higher sensitivity to the structural properties of the composite material allows to more sensitively follow the resin post-polymerization
- etc.

However, the low frequency usage was costly in the beginning of the 90th, which is not the case anymore and this off-line monitoring procedure can be easily implemented and could be interesting to complement on-line monitoring which is de facto happening at power frequency.

5. Conclusion

tan δ measurement. It reminds the theoretical bases before present two different cases: a coil and a stator. Measurement were performed on two frequencies: 50 Hz (power frequency) and 0.1 Hz.

If similar diagnostic results together with partial discharge measurements, different defects can be detected or identified at lower power frequency.

The low frequency measurements present some advantages. Especially the dielectric losses react more sensitive to structural changes.

However, while diagnostics can be improved using low frequencies, criteria must still be developed if the low frequency method should be used for acceptance testing. Indeed, most of the tan δ criteria are established for power frequency tests and we have to perform further research in order to define reliable acceptance criteria.

On the practical side, the instruments used by the authors have proven easily transportable, insensitive to noise and have very low power requirements in comparison to power frequency devices. So, that we feel, not only the enhanced diagnostic potential but also cost consideration, which justify more research efforts in that field.

References

[1] Christof Sumereder, Michael Muhr, Michael Marketz, Christian Rupp, Michael Kruger "Unconventionnal Diagnostic Methods for Testing Generator Stator Windings" IEEE Electrical Insulation Magazine , Vol. 25 No. 5, sept. 2009.

[2] A.K. Jonscher "Dielectric relaxation in solids" Chelsea Dielectric Press, London, 1983.

[3] Y. Feldmann, Y.A. Gusev, M.A. Vasilyeva 'Dielectric phenomena in complex systems" Tutorial Kazan Federal University, 2012.

[4] B. Fruth, G. Liptak, "Dielectric Properties of Mica Epoxy Composites Subjected to Thermal and Thermoelectrical Aging", Proc. 6th Int. Symp. High Voltage Engineering, New Orleans, 1989, paper 21.02

[5] IEEE 433-2009 : "IEEE Recommended Practice for Insulation Testing of AC Electric Machinery with High Voltage at Very Low Frequency".

[6] IEC 60034-27-3 "Rotating electrical machines – Part 27-3: Dielectric dissipation factor measurement on stator winding insulation of rotating electrical machine" 2015.

[7] C. Caironi "Contribution to the on-line forecast maintenance of electric machines by the analysis of signals related to physical phenomena relating thereto." PhD thesis, Université de Lorraine, France, 2002.

[8] C. Caironi, L. Durantay, D. Brie, A. Rezzoug "Interest & utility of Time-frequency and Time scale Transforms in the Partial Discharges Analysis" in proc. ISEI'2002, Boston USA.

[9] Nathaniel Taylor "Dielectric response and partial discharge measurements on stator insulation at varied low frequency", PhD thesis Stokholm, Sweden, 2010.

[10] B. Fruth, J. Fuhr “Partial discharges recognition – A Tool for diagnosis and Monitoring Ageing” In proc CIGRE’1990 Paris, France.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 160-167 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Analysis of Outdoor and Indoor Propagation at 15 GHz and Millimeter Wave Frequencies in Microcellular Environment

Muhammad Usman Sheikh*, Jukka Lempiainen

Tampere University of Technology, Department of Electronics and Communications Engineering, Finland.

ARTICLE INFO

Article history:
Received: 26 November, 2017
Accepted: 07 January, 2018
Online: 30 January, 2018

Keywords:
Multipath propagation
Microcellular
3D ray tracing
System performance
5G
Millimeter wave frequencies

ABSTRACT

The main target of this article is to perform the multidimensional analysis of multipath propagation in an indoor and outdoor environment at higher frequencies i.e. 15 GHz, 28 GHz and 60 GHz, using "sAGA" a 3D ray tracing tool. A real world outdoor Line of Sight (LOS) microcellular environment from the Yokusuka city of Japan is considered for the analysis. The simulation data acquired from the 3D ray tracing tool includes the received signal strength, power angular spectrum and the power delay profile. The different propagation mechanisms were closely analyzed. The simulation results show the difference of propagation in indoor and outdoor environment at higher frequencies and draw a special attention on the impact of diffuse scattering at 28 GHz and 60 GHz. In a simple outdoor microcellular environment with a valid LOS link between the transmitter and a receiver, the mean received signal at 28 GHz and 60 GHz was found around 5.7 dB and 13 dB inferior in comparison with signal level at 15 GHz. Whereas the difference in received signal levels at higher frequencies were further extended in an indoor environment due to higher building penetration loss. However, the propagation and penetration loss at higher frequency can be compensated by using the antenna with narrow beamwidth and larger gain.

1. Introduction

This article is an extension of research work originally presented at International Wireless Communication and Mobile Computing (IWCMC'17) conference [1]. In reference [1], studies were made at 15 GHz only, and only outdoor propagation was studied. However, in this article the radio propagation at millimeter wave frequencies i.e. 60 GHz is additionally studied. Furthermore, the research work of this article also includes indoor propagation analysis, which was not done earlier in [1].

Nowadays, the Fifth Generation (5G) of the mobile communication system is being actively discussed in both industry and academia [2-4]. Currently, various advanced wireless access technologies including High Speed Packet Access (HSPA), Long Term Evolution (LTE), and LTE-Advanced (LTE-A) are in operation. However, the ultimate solution for the ultra high capacity requirement of the future system is expected to be provided by the 5G. It is strongly believed that the higher frequency bands will play a vital role in meeting the capacity targets of the next generation of the cellular networks. The frequency bands between 700 MHz to 4 GHz are currently used by the mobile operators for radio transmission, and are already over loaded with current mobile technologies. Higher frequency bands offer wider spectrum. On the other hand, the higher path loss is also attributed with the higher frequency of operation [5]. The R&D organizations are extensively putting their effort for investigating the utilization of higher frequency bands for mobile communications [6]. The utilization of advanced antenna technologies such as Massive MIMO (MMIMO) and beamforming help in extending the coverage at higher frequencies.

It is important to understand the radio propagation characteristics in order to properly design and to efficiently optimize the system. Classical coverage prediction models do- not provide the insight information about the channel (environment). However, the deterministic ray tracing models provide the multidimensional characteristics of the radio propagation environment [7]. Multipath propagation is a complex phenomenon and involves several propagation mechanisms e.g. specular

*Corresponding Author : Muhammad Usman Sheikh,
Email : muhammad.sheikh@tut.fi

reflection, diffraction from the corners, transmission, diffraction from the rooftops, and diffused scattering [8]. Ray tracing is a promising technique for finding the possible paths between the transmitter and receiver.

At reference [9], the NTT DOCOMO provides the field measurement results of 5G radio system operating at the frequency of 15 GHz in an outdoor and indoor microcellular environment. One of the core targets of this article is to provide the simulation result for the scenario considered in [9], so that a comparison between the measured and simulated data can be made. For this purpose, the sAGA ray tracing tool based on Image Theory (IT) algorithm is utilized for the simulations. The radio propagation properties such as received signal strength, Power Angular Spectrum (PAS), and Power Delay Profile (PDP) at higher frequencies are also provided in this article. The simulation results presented in this article highlight the difference of propagation at different considered frequencies i.e. 15 GHz, 28 GHz and 60 GHz, and also highlight the difference of propagation in an outdoor and indoor environment.

2. System Model and Environment

2.1. System Configuration

The Fifth Generation (5G) of the mobile communication system based on Time Division Duplex (TDD) operating at 15 GHz frequency is considered at reference [9]. The 5G communication system consists of four contiguous Component Carriers (CCs), and each CC is assumed to have 100 MHz bandwidth. Carrier Aggregation (CA) is employed to combine four component carriers. The transmission power per component carrier is set to 27.3 dBm, which makes a total transmission power of the transmitter equals 33.32 dBm for 400 MHz bandwidth. The base station is installed with a directional antenna at the height of 8 m. The transmitter is located on the wall of the building. The base station antenna has a horizontal Half Power Beamwidth (HPBW) of 90° and a vertical HPBW of 10.5° and has a maximum antenna gain of about 14.5 dBi. A mechanical downtilt of 13.5° is used to restrict the propagation in small cell environment. Whereas at the receiver end the MS is assumed to have an omni directional antenna with 0 dBi gain. As, the target of this paper was to provide the simulation results for the system and environment considered in [9], therefore the same set of parameters and environment are assumed in this paper.

2.2. Simulation Platform

The "sAGA" a MATLAB based three-dimensional ray tracing tool is used for the simulations. The sAGA tool is indigenously developed by the authors of this paper. Unlike other quasi 3D ray tracing tool, sAGA performs full three dimensional ray tracing. Multipath propagation involves several mechanisms of interactions e.g. reflection, diffraction, transmission, and scattering. The sAGA ray tracing tool uses Image Theory (IT) to find all the possible reflected, diffracted, ground reflected, and rooftop diffracted paths with the given number of reflections and diffractions [10-12]. In case of reflected paths, the loss in energy due to specular reflection is given by the reflection co-efficient and depends upon the incident and the reflected angle of path and material permittivity. Whereas for the diffracted paths the diffraction loss is given by Berg's recursive model [13]. The energy is scattered in the wide range of directions and the impact of scattering becomes significant at higher frequencies. Therefore, a concentric circle approach given at [14] is used to generate the scattering points on the walls of the building. Ray tracing requires detailed information about the simulation environment and provides accurate results. Three dimensional ray tracing requires 3D map of the simulation environment. The computational complexity of the ray tracing algorithm increases with the increase in the number of supported reflections and diffractions.

A. Simulation Environment

For simulations, a small area from the Yokusuka city of Japan is selected. The Google map view of the targeted area is shown in Figure 1. For outdoor microcellular environment simulation, the parking area of NTT DOCOMO R&D center in Japan is considered, and for indoor simulation the office building is considered in this article.

Figure 1. Google map view of targeted area.

Figure 2 illustrates the two dimensional simplified map of the considered area. The simplified map model consists of just five buildings. For outdoor coverage simulations, it is assumed that the MS is mounted on a vehicle at a height of 3.1 m. The MS is driven at the speed of 10 km/h along eight different routes (A1-A8) as shown in Figure 2. For indoor coverage simulations, two paths A9 and A10 are considered. The path A9 is close to the exterior wall of the building, and path A10 is 7 m away from the exterior wall of the building. The location of the transmitter is marked with the blue spot. The transmitter antenna has the azimuth angle of 90° (facing towards North). All the outdoor and indoor simulation routes have a clear LOS with the transmitter. Two outdoor location points are marked as Pt1 and Pt2, and one indoor location point is marked as Pt3 in Figure 2. These static points are selected for Power Angular Spectrum (PAS) analysis.

For PAS simulations, a directive antenna with 14° HPBW in the horizontal domain and 10.5° HPBW in the vertical domain with 20 dBi maximum gain is used. It is assumed that a directive antenna on a receiver side at a height of 1.65 m is rotated by 360° in the azimuth plane with a step size of 4°. For ray tracing, the reflected path with the maximum three reflections, and diffracted path with single diffraction were found. The scatterers are spread over the surface of the walls of the building. The general system simulation parameters are summarized in Table I.

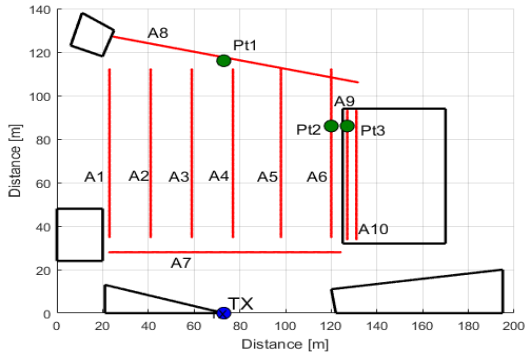


Figure 2. Illustration of simulation area and simulation routes.

Table I. General system simulation parameters for route simulations.

Parameters	Unit	Value
Frequency	GHz	15/28/60
Single carrier component (CC) bandwidth	MHz	100
System bandwidth (4 CCs)	MHz	400
Transmission power per CC	dBm	27.3
Total transmission power	dBm	33.32
Transmitter height	m	8
Antenna downtilt	°	13.5
MS height	m	3.1
Reflections		3
Diffractions		1
Diffuse scattering		Enabled

2.3. Building Penetration Loss (BPL)

The signal experiences a penetration loss while penetrating from the outdoor environment to the indoor environment. Outdoor to indoor penetration loss is generally termed as Building Penetration Loss (BPL). The building penetration loss is the function of frequency and depends heavily on the material characteristics of the building; therefore the BPL can be significantly different for different material types at different frequencies. Generally, the old houses are composed of plane standard glass windows and concrete wall, while the Infrared Reflective (IRR) glass windows are commonly used in the new modern energy saving houses. In reference [15], the old buildings are assumed to have 30 % of the standard glass windows and 70 % of the concrete wall. Similarly, the assumption for new modern building type corresponds to the 70 % of the IRR glass windows and 30 % of the concrete wall.

A simple model structure has been proposed in [15] to model a single material frequency dependent penetration loss. The penetration loss for different material types is provided at [16-20].

$$L_{\text{Single glass,dB}} = 0.1 * \text{Frequency}_{\text{GHz}} + 1, \quad (1)$$

$$L_{\text{Double glass,dB}} = 0.2 * \text{Frequency}_{\text{GHz}} + 2, \quad (2)$$

$$L_{\text{IRR glass,dB}} = 0.3 * \text{Frequency}_{\text{GHz}} + 23, \quad (3)$$

The penetration loss for the concrete wall as a function of frequency is modeled as

$$L_{\text{Concrete,dB}} = 4 * \text{Frequency}_{\text{GHz}} + 5, \quad (4)$$

As the buildings are composite of windows and concrete wall, the building penetration loss for old buildings and new buildings is modeled as shown in (5) and (6), respectively [21].

$$L_{\text{Old building,dB}} = -10\text{Log}_{10}\left[0.3 * 10^{\frac{-L_{\text{Double glass,dB}}}{10}} + 0.7 * 10^{\frac{-L_{\text{Concrete,dB}}}{10}}\right], \quad (5)$$

$$L_{\text{New building,dB}} = -10\text{Log}_{10}\left[0.7 * 10^{\frac{-L_{\text{IRR glass,dB}}}{10}} + 0.3 * 10^{\frac{-L_{\text{Concrete,dB}}}{10}}\right], \quad (6)$$

The building penetration loss as a function of frequency for different building types is shown in Figure 3.

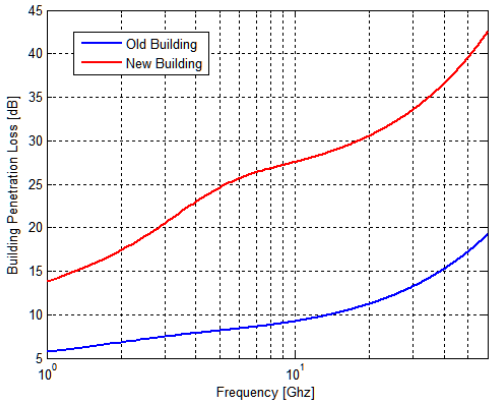


Figure 3. Building penetration loss as a function of frequency.

2.4. Indoor Propagation Loss

In an indoor environment, generally the indoor walls are made up of standard glass alternatively plaster. In [15], two different indoor wall loss models are presented as a function of the frequency assuming an average wall distance of 4 m. The Indoor Loss Model 1 assumes an indoor wall of standard glass, whereas Indoor Loss Model 2 is based on the measurements performed in [16]. Two indoor wall loss models are modeled as shown in (7) and (8). Indoor loss as a function of frequency for two different indoor wall loss models, expressed as dB/m is shown in Figure 4.

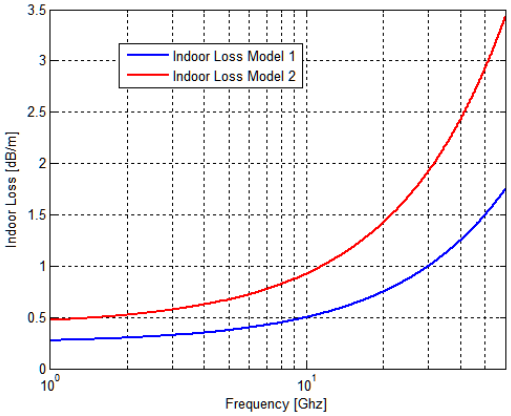


Figure 4. Indoor loss models as a function of frequency.

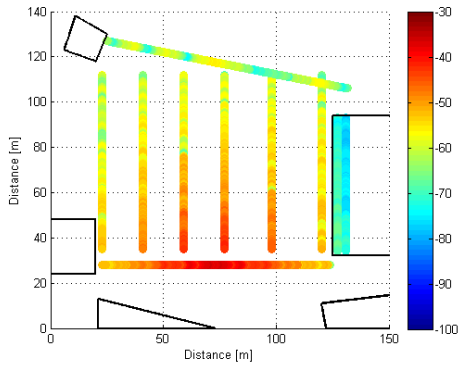
$$L^{(1)}_{\text{Wall loss,dB/m}} = L_{\text{Single glass,dB}}, \quad (7)$$

$$L^{(2)}_{\text{Wall loss,dB/m}} = 0.2 * \text{Frequency}_{\text{GHz}} + 1.7, \quad (8)$$

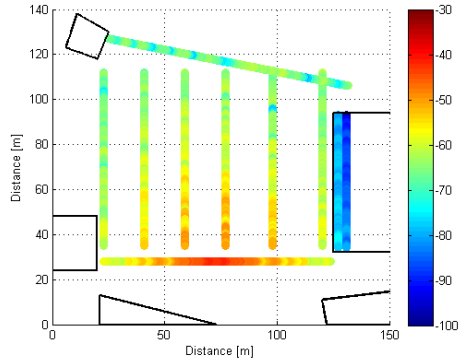
3. Simulation Results and Discussions

Figure 5 shows the received signal strength in dBm for eight outdoor and 2 indoor simulation routes. Figure 5(a), Figure 5(b) and Figure 5(c) shows the received signal strength along the paths at 15 GHz, 28 GHz and 60 GHz, respectively. It can be seen in Figure 5(a) that the maximum received signal strength of around -38 dBm is found in front of the BS antenna in the direction of the main lobe. However, the signal strength starts to degrade as the receiver starts to move away from the transmitter, and deviate from the direction of the main lobe of an antenna. The considered area for simulation was an open environment; therefore the LOS path always existed between the transmitter and receiver. Whereas there are signal fades along the routes due to the constructive and destructive addition of multipaths. The route A1 and A6, and path A2 and A5 are symmetrical and are at almost equal distance from the transmitter. It is interesting to see that the signal strength along the route A6 is better compared with A1 due to more reflected and scattered multipaths from the wall of the nearby building. As soon as the MS crossed the building in route A6, the signal degradation due to the absence of a reflected path is witnessed. The mean received signal level of path A6 is -57.01 dBm at 15 GHz. For indoor paths, old building type is considered. Path A9 is adjacent to path A6. However, due to building penetration loss the mean received signal level for path A9 is -69.78 dBm assuming old building type. The mean received signal level is further drops to -75.12 dBm in case of new building type which consists of IRR glass. Path A10 is located more deep inside the building, and due to additional indoor wall loss the signal strength degrades more rapidly in an indoor environment. The difference between the received signal level at 15 GHz and two other higher frequencies is clearly evident. However, the path loss at higher frequency can be compensated by using the antenna with higher gain.

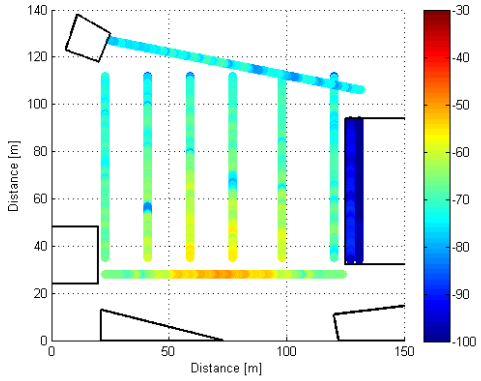
Table II presents the mean received signal level for the different simulation routes at 15 GHz, 28 GHz and 60 GHz. Considering the outdoor simulation routes, the mean path loss difference of around 5.7 dB and 13 dB was found between 15 GHz and 28 GHz transmission, and between 15 GHz and 60 GHz transmission, respectively. Similarly, for indoor simulation paths the mean path loss difference is around 10.57 dB and 27.82 dB between 15 GHz and 28 GHz transmission, and 15 GHz and 60 GHz transmission, respectively. As stated earlier, that the simulation routes A1 and A6 are symmetrical, however due to the presence of nearby wall for route A6 the mean received signal strength is around 1.25 dB and 1.43 dB better than A1 at 15 GHz and 28 GHz, respectively. Building penetration loss is the function of frequency and therefore the signal experiences higher penetration loss at higher frequencies. A significant difference was found between the received signal level of indoor user located in an old and new building type.



(a)



(b)



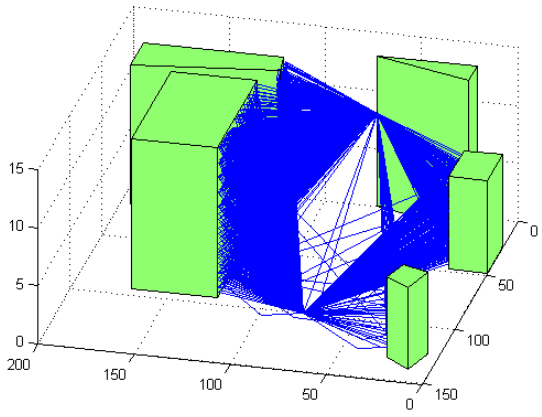
(c)

Figure 5. Received signal strength along the simulation route at, (a) 15 GHz, (b) 28 GHz, and (c) 60 GHz.

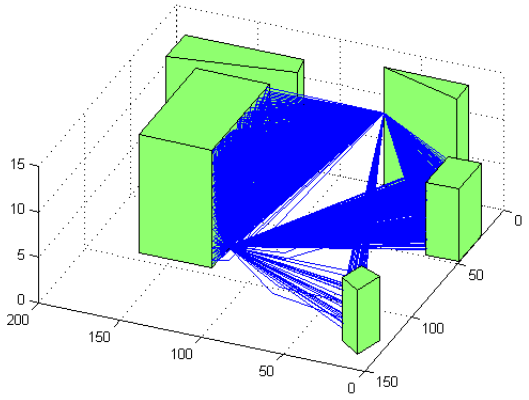
Table II. Mean RX level.

Simulation route	Mean RX level (dBm)		
	15 GHz	28 GHz	60 GHz
A1	-58.26	-64.16	-70.15
A2	-54.96	-60.61	-68.57
A3	-53.23	-58.46	-65.45
A4	-52.37	-58.32	-66.45
A5	-53.92	-60.88	-66.79
A6	-57.01	-62.73	-70.98
A7	-45.54	-51.49	-59.51
A8	-63.09	-67.17	-74.90
A9 (Old building)	-69.78	-78.54	-91.65
A10 (Old building)	-75.12	-86.20	-104.4
A9 (New building)	-88.61	-98.66	-114.97
A10 (New building)	-93.94	-106.33	-127.73

Figure 6(a) and Figure 6(b) show the cluster of rays at 15 GHz reaching the outdoor location Pt1 and Pt2, respectively.



(a)

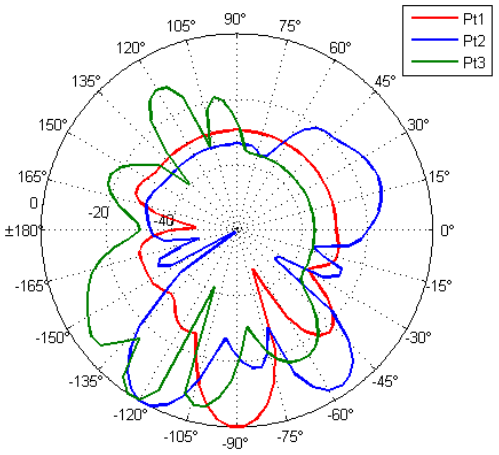


(b)

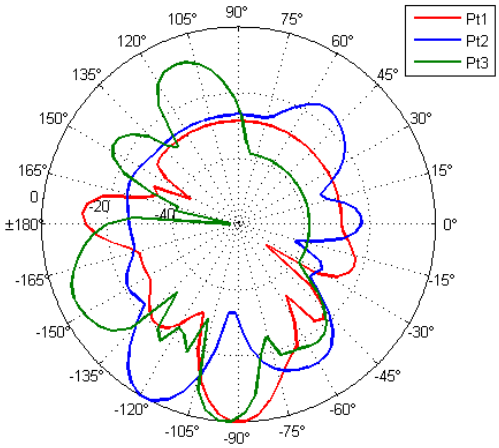
Figure 6. Multipath rays at 15 GHz from TX to (a) Pt1, and (b) Pt2.

The cluster of rays shown in Figure 6 includes wall reflected, ground reflected, diffracted and scattered paths. The impact of diffused scattering is clearly evident, and a considerable amount of energy at receiver points is coming through a large number of scattered paths. In Figure 6(a), the Pt1 is located in the middle of the simulation environment; therefore the walls of the building have clear visibility to the Pt1. Whereas, Pt2 is located closed to the wall of the nearby building, and the building on the right side of the TX does not have visibility to Pt2. Therefore, it can be seen in Figure 6(b) that there is no scattered path from the building on the right side of the transmitter.

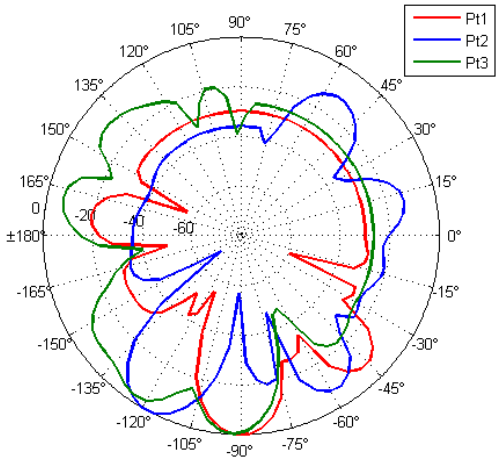
Figure 7(a), Figure 7(b) and Figure 7(c) shows the normalized power angular spectrum for location Pt1, Pt2, and Pt3 at 15 GHz, 28 GHz, and 60 GHz, respectively. The PAS shown in Figure 7 confirms that in the LOS condition the major contribution to the total received power comes through the LOS path, and through the specular reflected paths. However, at 28 GHz and 60 GHz the impact of diffused scattering is more significant than 15 GHz. An indoor environment has more scattering components due to nearby walls. It can be seen in Figure 7 that phenomenon of scattering has more significance at higher frequencies specifically in an indoor environment. The Pt1 is located at an azimuth angle of 90°, and therefore at Pt1 the strongest LOS path has a Direction of Arrival (DoA) of -90° as shown in Figure 7. The next strongest path is the reflected path with -44° of DoA.



(a)



(b)



(c)

Figure 7. Power angular spectrum for location Pt1, Pt2 and Pt3, (a) at 15 GHz, and (b) at 28 GHz.

The LOS path has a 3D path length of 116 m; whereas the reflected path has a path length of 156 m i.e. 40 m longer than the LOS path. Therefore, the strength of the reflected path is almost 17 dB and 22 dB less compared with the LOS path at 15 GHz and

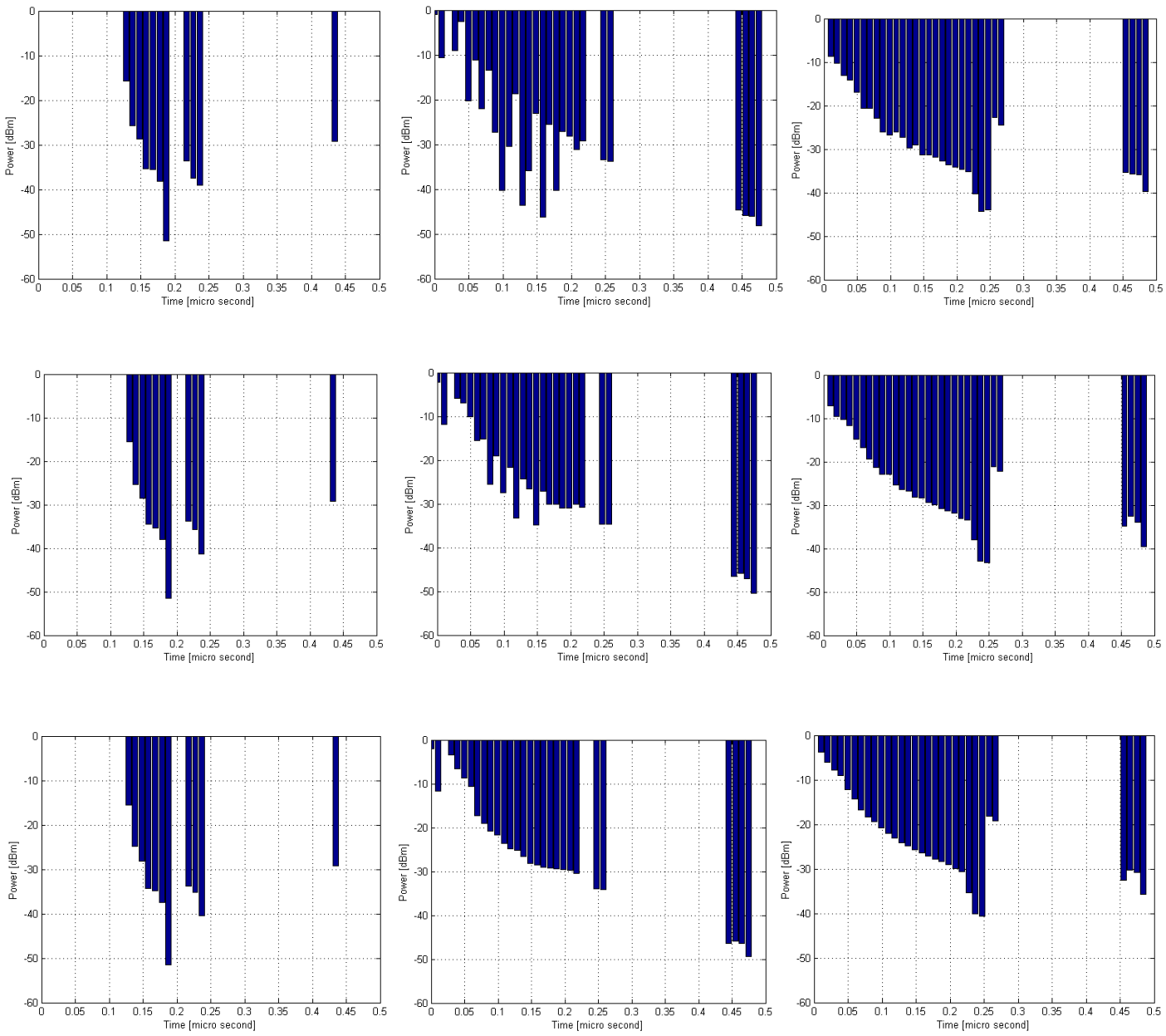


Figure 8. Power delay profile, (a) Pt1 (15GHz), (b) Pt2 (15GHz), (c) Pt3 (15GHz), (d) Pt1 (28GHz), (e) Pt2 (28GHz), (f) Pt3 (28GHz), (g) Pt1 (60GHz), (h) Pt2 (60GHz), and (i) Pt3 (60GHz).

28 GHz, respectively. However, at 60 GHz due to the presence of a large number of scattering components the composite strength of received signal components at -44° of DoA is almost 10 dB less compared with LOS direction. The Pt2 is located on the right side of the transmitter and has a valid LOS link with the transmitter

In Figure 7, the strongest path (LOS path) has around -120° of DoA and the other reflected path is reaching the receiver at -60°. At Pt2, the direct path has a path length of 98 m, and the reflected path has a path length of 104 m i.e. 6 m longer than the LOS path. Therefore, the strength of the reflected path is around 4 dB and 11 dB less compared with the LOS path at 15 GHz and 28 GHz, respectively. Also, the impact of diffused scattering from the wall at an angle of -60 ° at 28 GHz and 60 GHz is visible. The considered microcellular case is a simple environment, and not a multipath rich outdoor environment. Therefore, the most of the energy at the receiver points is coming from a narrow direction only. However, in an indoor environment, due to a rich indoor scattering environment the PAS of incoming multipath is wide spread, especially at higher frequencies i.e. 28 GHz and 60 GHz.

Figure 8 shows the Power Delay Profile (PDP) acquired through simulations of location Pt1, Pt2 and Pt3 at 15 GHz, 28 GHz and 60 GHz. For Pt1, the reflected path has 40 m longer path length therefore the reflected path reaches the receiver point with a delay of around 0.125 μs relative to LOS path, and the signal strength of the reflected path is around 16 dB less compared with LOS path. For Pt1, the power delay profile is almost identical at

15 GHz, 28 GHz, and 60 GHz. The Pt2 is located close to the wall of the building, and therefore the difference of path length between the reflected path and the LOS path is short. By comparing Figure 8(b), Figure 8(e), and Figure 8(h), it can be noted that the reflected path with slightly less signal strength than LOS path reaches the receiver point within a very short delay of 0.01 μs. Other diffracted and scattered paths from the nearby wall also reach the receiver Pt2 with a short delay period. The power delay profile of Pt2 got smoother at higher frequencies due to a large number of scattering paths at higher frequencies.

Indoor location Pt3 is located nearby the outdoor location Pt2. However, an indoor environment is a scattering rich environment and a larger number of scattering paths reach the receiver point with small delay. Therefore, the PDP of Pt3 is smoother compared with PDP of Pt2. It is interesting to find that the phenomenon of scattering becomes more prominent in higher frequencies, and the power of multipath components with short delay improves at 28 GHz and 60 GHz. The mean delay spread in nanoseconds for static location points are given in Table III.

Table III. Mean delay spread.

Location	Mean Delay Spread (ns)		
	15 GHz	28 GHz	60 GHz
Pt1	150	150	150
Pt2	109	107	106
Pt3	124	119	114

4. Conclusion

In this article, a multidimensional analysis of the multipath propagation in an outdoor and indoor microcellular environment has been carried out at higher frequencies i.e. 15 GHz, 28 GHz and 60 GHz, by means of a sophisticated ray tracing tool. The simulations are performed using a simplified 3D map of a small area from the Yokusuka city of Japan. The considered outdoor simulation routes and outdoor static points represent an open environment scenario with valid LOS link between the transmitter and the receiver in a microcellular environment. Indoor simulation routes and an indoor static point were also considered for studying indoor propagation at higher frequencies. The simulation results presented for different simulation routes at 15 GHz are in a close relationship with a measurement results presented in [9]. Similarly at 15 GHz for the selected static points the power angular spectrum (PAS) at receiver end acquired through simulations are in a satisfactory agreement with the measured PAS given at [9]. The power angular spectrum (PAS) at receiver shows the direction of arrival of LOS, specular reflected, diffracted and diffused scattered paths. It was found that mean received signal strength in an outdoor environment for the selected simulation routes at 15 GHz is almost 5.7 dB, and at 60 GHz the mean RX level is almost 13 dB inferior in comparison with the propagation at 15 GHz due to higher propagation loss at higher frequencies. The difference in a mean received signal in an indoor environment between 15 GHz and 28 GHz is extended to nearly 10. 5 dB, and the difference between the mean received signal in an indoor environment between 15 GHz and 60 GHz is further stretched to 28.5 dB due to extra building penetration loss at higher frequencies. However, the higher path loss at 28 GHz and 60 GHz can be compensated by using an antenna with higher gain. The power delay profile acquired through simulations was found helpful in distinguishing the multipath richness of the environment. The indoor was found more scattering rich environment compared with outdoor environment. The absolute values of performance results presented in this article may vary depending upon the modeling impairment, antenna radiation pattern, simulation parameters and simulation environment.

Acknowledgment

Authors would like to thank European Communications Engineering (ECE) Ltd for supporting this research work.

References

[1] M. U. Sheikh and J. Lempiäinen, "Analysis of multipath propagation for 5G system at higher frequencies in microcellular environment," 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, 2017, pp. 1660-1664.

[2] NTT DoCoMo, Inc, "5G Radio Access: Requirements, Concept and Technologies," White Paper, 2014.

[3] Huawei Technologies, "5G: A Technology Vision", White Paper, 2013.

[4] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Selén, "5G radio access," Ericsson Review, vol. 91, no. 6, pp. 42–48, 2014.

[5] T. S. Rappaport , J. N. Murdock and F. Gutierrez , "State of the art in 60 GHz integrated circuits & systems for wireless communications" , Proc. IEEE , vol. 99 , pp.1390 -1436 , 2011.

[6] S. G. Larew, T. A. Thomas, M. Cudak, A. Ghosh, "Air Interface Design and Ray Tracing Study for 5G Millimeter Wave Communications" in Proc. IEEE Globecom 2013, Atlanta, USA, 9-13 Dec, 2013.

[7] F. Fuschini, H. El-Sallabi, V. Degli-Esposti, L. Vuokko, D. Guiducci and P. Vainikainen, "Analysis of Multipath Propagation in Urban Environment Through Multidimensional Measurements and Advanced Ray Tracing Simulation," in IEEE Transactions on Antennas and Propagation, vol. 56, no. 3, pp. 848-857, March 2008.

[8] T. S. Rappaport, "Wireless Communications: Principles and Practice", 2nd Edition, 2001.

[9] K. Tateishi et al., "Field experiments on 5G radio access using 15-GHz band in outdoor small cell environment," Personal, Indoor, and Mobile Radio Communications (PIMRC), 2015 IEEE 26th Annual International Symposium on, Hong Kong, 2015, pp. 851-855.

[10] H. W. Son, and N. H. Myung, "A Deterministic Ray Tube Method for Microcellular Wave Propagation Prediction Model," in IEEE Transactions on Antennas and Propagation, vol. 47, no. 8, pp. 1344-1350, August 1999.

[11] S. Soni, and A. Bhattacharya, "An Efficient Two-dimensional Ray-tracing Algorithm for Modeling of Urban Microcellular Environment", in International Journal of Electronics and Communications, vol. 66, no. 6, pp. 439-447, June 2012.

[12] D. N. Schettino, F. J. S. Moreira and C. G. Rego, "Efficient Ray Tracing for Radio Channel Characterization of Urban Scenarios," in 12th Biennial IEEE Conference on Electromagnetic Field Computation, Miami, FL, 2006, pp. 267-267.

[13] J.-E. Berg, "A Recursive Method for Street Microcell Pathloss Calculations", Sixth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC'95, Wireless: Merging onto the Information Superhighway. Vol. 1, 1995.

[14] X. Li, "Efficient Ray Tracing Simulation", Master's thesis at LUND University, 2014.

[15] E. Semaan, F. Harrysson, A. Furuskär and H. Asplund, "Outdoor-to-indoor coverage in high frequency bands", 2014 IEEE Globecom Workshops (GC Wkshps), Austin, TX, 2014, pp. 393−398.

[16] C. Larsson, F. Harrysson, B.-E. Olsson, and J. E. Berg, "An outdoor-to-indoor propagation scenario at 28 GHz," in 8th European Conference on Antennas and Propagation (EuCAP 2014), The Hague,The Netherlands, April 2014, pp. 3301−3304.

[17] W. C. Stone, “Electromagnetic signal attenuation in construction materials, ”NIST Building and Fire Research Laboratory, Gaithersburg, Maryland, NISTIR 6055 Report No. 3 6055, Oct. 1997.

[18] L. M. Frazier, “Radar surveillance through solid materials,” in Proceedings of the SPIE - The International Society for Optical Engineering, vol. 2938, Hughes Missile Syst. Co., Rancho Cucamonga, CA, USA, 1997, pp. 139−146.

[19] R. Wilson, “Propagation losses through common building materials 2.4 GHz vs 5 GHz,” University of Southern California, CA, Tech. Rep. E10589, Aug. 2002.

[20] C. A. Remley, G. H. Koepke, C. L. Holloway, C. A. Grosvenor, D. G. Camell, J. M. Ladbury, R. Johnk, D. R. Novotny, W. F. Young, G. Hough, M. McKinley, Y. Becquet, and J. Korsnes, “Measurements to support modulated-signal radio transmissions for the public-safety sector,” NIST, Boulder, CO, Tech. Rep. Tech. Note 1546, Apr. 2008.

[21] White paper on “5G Channel Model for bands upto 100 GHz”.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 168-190 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Linear algebra as an alternative approach to the synthesis of digital devices of automation and control systems

Nikolay Chernov[1], Nikolay Prokopenko[*,2,3], Vladislav Yugai[1], Nikolay Butyrlagin[2]

[1]*Systems of automation control, Southern Federal University, SFedU, Taganrog, 347928, Russia*

[2]*Information systems and radioengineering, Don State Technical University, DSTU, Rostov-on-Don, 344000, Russia*

[3]*Institute for Design Problems in Microelectronics of Russian Academy of Sciences, IPPM RAS, Zelenograd, 124681, Russia*

ARTICLE INFO

Article history:
Received: 18 January, 2018
Accepted: 18 January, 2018
Online: 30 January, 2018

Keywords :
multivalued component base
multivalued logic
linear logic synthesis

ABSTRACT

The article considers linear algebra as an alternative mathematical tool of logic synthesis of digital structures to Boolean algebra and synthesis methods of digital electronic component base (ECB) on its ground. The methods of solving the applied problems of logic synthesis are shown, including the expansion of an arbitrary logic function by means of monotonic functions. The proposed mathematical apparatus actually provides the creation of digital structures on the principles of analog circuitry. It can find application in the design of multivalued digital ECB, specialized system-on-chip and analog-digital sensors with current output. The examples of synthesis of the combinational and sequential two-valued and multivalued digital devices are given. In conclusion, the advantages of linear algebra in comparison with Boolean algebra are formulated.

1. Introduction

This article is a continuation of the studies presented at the conference SIBCON-2017 [1].

Boolean algebra is known [2] as a leading mathematical tool for logic synthesis of two-valued digital structures. Almost all existing methods of logic synthesis are formed on its ground. The success of Boolean algebra is caused, among other things, by the fact that the Boolean representation of the realized logic function turned out to be rather technological: the circuit implementation of logic elements was relatively simple. TTL, C-MOS and ESL and other technologies occurred to be the most preferable for this purpose.

Despite this, the history of the development of digital microelectronics knows the attempts to replace both the approach to logic synthesis (the use of spectral representations [3 - 6] using the arithmetic polynomials [7 - 9]) and technological realization (I^2L, I^3L, ...).

The most significant contribution to the alternative theory of logic synthesis has been made by the threshold interpretation of Boolean algebra, called the threshold logic.

*Corresponding Author: Nikolay Prokopenko, Email: prokopenko@sssu.ru

Threshold logic related to one of the directions of synthesis of the digital structures [10], is constantly evolving. We know a significant number of publications devoted to this problem [11 – 14]. For two-valued functions, the threshold synthesis problems have already been solved beginning from the well-known papers of M. Dertouzos [10]. One of the advanced approaches of synthesis for the *k*-valued threshold functions is considered in this article [11].

The threshold logic was initially implemented in a two-valued version, but many publications on the multivalued threshold logic have recently appeared [14 - 17]. Why are they dedicated to the multivalued logic? The fact is that the multivalued logic is currently considered as a means of improving the quality characteristics of LSI (the ratio of "number of elements / number of links", "total area / real-estate" of the crystal, etc.), which doesn't require a drastic change in LSI technology. The authors of this article also follow this approach.

To develop this ideology we proposed an alternative approach to the logic synthesis of digital devices - the replacement of the mathematical tool of Boolean algebra by another body of mathematics - linear algebra. Such a replacement entailed fundamental changes in various aspects of the implementation of digital structures:

- the potential realization of Boolean logic was replaced by the current linear implementation;

- the key operating mode of the elements was replaced by the linear one;

- Boolean values of the variables took a quantitative form instead of the qualitative (logic) one;

- the value of logic was determined not by the scheme, but by the significance of signals;

- the output signal was presented by the difference of signals of two circuits operating in parallel, which improves the performance of digital structures.

Thus, the use of linear algebra as a mathematical tool of the logic synthesis ensured the development of alternative methods for the logic synthesis of current digital circuits and their practical implementations.

The authors published a number of papers [18-32] devoted to the synthesis of logic (nonthreshold) two-valued and multivalued digital structures. This paper considers the use of linear algebra as a mathematical tool for the logic synthesis of two-valued and multivalued, logic and threshold digital structures.

The fuzzy concept of the "threshold synthesis" can be interpreted in two ways:

- as a normal logic synthesis of digital structures with circuit implementation on threshold logic elements (any logic function can be implemented in this way);

- as a logic synthesis of threshold logic functions (an arbitrary logic function can be implemented by a network of threshold logic elements).

The purpose of this article is to propose logical and threshold current hardware components for constructing digital structures within the two specified areas. Within the framework of this goal, the authors' solutions for the following tasks are given below:

- since linear algebra is used as a mathematical tool for the logic synthesis of current logic structures, the article presents the main definitions and aspects of the practical use of linear algebra;

- as there is a close relationship between the threshold and monotonic functions, the article gives a definition of the monotonic function and explains the ways of representing arbitrary functions by monotonic linear functions;

- some versions of transformation of two-valued and multivalued monotonic functions into a threshold form in linear algebra are analyzed;

- versions of circuit implementation of linear threshold logic elements are considered.

2. Linear Algebra

A. Definition of linear algebra. Let $\boldsymbol{P} \to \langle P; +, \cdot, 0.1\rangle$ – field, $\langle A; +, \cdot, \theta\rangle$ – algebra with two binary operations and one nullary operation. The system $\boldsymbol{\Lambda} \to \langle A; +, \cdot, \theta, \mathbf{P}\rangle$ is called *linear algebra*, if the following conditions are met:

– the system $\boldsymbol{\Lambda} \to \langle A; +, \cdot, \theta, \boldsymbol{P}\rangle$ – linear (vector) space over the field $\boldsymbol{P}$;

– distributivities of operations + and ·

$$\forall(a, b, c \in A)\ (a + b)c = ac + bc \wedge c(a + b) = ca + cb;$$

– associativities of vector multiplication by elements of the field $\boldsymbol{P}$

$$\forall(a, b \in A) \wedge \forall(k \in P)\ k(ab) = (ka)b = a(kb).$$

B. Extension of linear algebra.

Let $\boldsymbol{A} \to \langle A; +, \cdot, \theta\rangle$ – vector space of linear algebra $\boldsymbol{\Lambda}$, $\boldsymbol{P} \to \langle P; \Omega = \{\omega_k | k \in P\}, 0.1\rangle$ – field of linear algebra $\boldsymbol{\Lambda}$, which contains the operations ω_k, which in general case do not necessarily coincide with the operations of linear space $\boldsymbol{A}$. Then the system $\boldsymbol{\Lambda'} \to \langle \{A; +, \cdot, \theta\}, \{P; \Omega, 0, 1\}\rangle$ is called *the extension of linear algebra* $\boldsymbol{\Lambda}$.

When interpreting this algebraic system in a certain way, we can obtain algebras with different properties. For example, interpreting A as a set of terms of Boolean functions $f(x_1,\ldots, x_n)$, the operations "+" and "·" - as $max(x_1,\ldots, x_n)$, $min(x_1,\ldots, x_n)$, we obtain algebra $\boldsymbol{A} \to \langle A; max, min, \theta; \boldsymbol{P}\rangle$. Similarly, leaving the semantics of operations in the original form (i.e., defining them as ordinary arithmetic operations), we can consider the reduced system as linear algebra on the set A of vectors in a linear space. The reduced algebraic system is considered below in this form.

C. Creation of bases. To form the bases from logic variables, it is possible to construct different constructions of linearly independent vectors with given properties. The choice of operations for creation of the bases is made independently on the operations of linear space and can be determined by various (mathematical, circuit, technological and other) requirements. In the applied plan, this allows obtaining the ideologically unified (based on operations of linear space) circuit solutions of functional elements (from different implementations based on field operations).

A basis is a system of m linearly independent vectors $\{F\} = \{\varphi_0, \varphi_1,\ldots, \varphi_{m-1}\}$, which enables to describe any vector $a \in \boldsymbol{A}$ in the linear form

$$a = \sum_{i=0}^{m-1} a_i \varphi_i\ , \qquad (1)$$

with respect to these vectors.

Each vector of the basis $\{F\} = \{\varphi_0, \varphi_1,\ldots, \varphi_{m-1}\}$ is uniquely determined by the set of coordinates $\varphi_i = \{\varphi_{i0}\ , \varphi_{i1},\ldots, \varphi_{i,m-1}\}$, with the help of which we can make a square matrix of order m:

$$F = \begin{bmatrix} \varphi_{10} & \varphi_{11} & \cdots & \varphi_{1,m-1} \\ \varphi_{20} & \varphi_{21} & \cdots & \varphi_{2,m-1} \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_{m-1,0} & \varphi_{m-1,1} & \cdots & \varphi_{m-1,m-1} \end{bmatrix},$$

corresponding to the given basis {F}.

Two bases $\{F\} = \{\varphi_1, \varphi_2, \ldots, \varphi_{m-1}\}$ and $\{Y\} = \{\gamma_1, \gamma_2, \ldots, \gamma_{m-1}\}$, the matrices F and Y of which are inverse to each other, are reciprocal (dual). Besides,

$$F \cdot Y = E,$$

where E – diagonal matrix of order m with $\varepsilon_{ij} = 1$, when $i = j$ and $\varepsilon_{ij} = 0$, if $i \neq j$, which is an orthonormal basis $\{E\}$.

Since the resolution of an arbitrary vector a over the basis $\{F\}$ has the form

$$a = a_0\varphi_0 + a_1\,\varphi_1 + \cdots + a_{m-1}\,\varphi_{m-1},$$

then, multiplying both parts of this resolution by γ_i, $i = [0, m-1]$, we obtain:

$$a_i = a\cdot\gamma_i = a\cdot(Y\cdot\varepsilon_i) = a\cdot(F^{-1}\cdot\varepsilon_i).$$

Then the procedure for representing (i.e. obtaining the values of the expansion coefficients) of the arbitrary vector a in the given basis {F} is reduced to performing the following operations:

– construction of the basic matrix F;

– construction of the inverse basic matrix F^{-1};

– multiplication of the row-vector a by the column-vector of the matrix F^{-1} and computation of the expansion coefficients of the vector a over the basis $\{F\}$;

– writing of the expression for the vector a in the linear form (1) with respect to the basis $\{F\}$.

Example 1. Get a representation of the conjunction operation of two arguments $x_1\&x_2$ of the value 2 in the basis

$$F = \begin{bmatrix}1 & 1 & 1 & 1\\ 0 & 0 & 1 & 1\\ 0 & 1 & 0 & 1\\ 0 & 1 & 1 & 0\end{bmatrix},\quad (F)^{-1} = \frac{1}{2}\begin{bmatrix}2 & -1 & -1 & -1\\ 0 & 1 & -1 & 1\\ 0 & -1 & 1 & 1\\ 0 & 1 & 1 & -1\end{bmatrix}.$$

Solution.

a) we represent the sequence of values of the two-valued logic function by the row-vector

$$x_1\&x_2 = [0, 0, 0, 1];$$

b) we multiply the resulting row-vector by the columns of the inverse basic matrix $(F)^{-1}$ and obtain the row-vectors of the expansion coefficients of the represented logic function with respect to the given basis and the expression of the conjunction operation

$$x_1\&x_2 = [0, 0, 0, 1]\cdot\frac{1}{2}\begin{bmatrix}2 & -1 & -1 & -1\\ 0 & 1 & -1 & 1\\ 0 & -1 & 1 & 1\\ 0 & 1 & 1 & -1\end{bmatrix} =$$

$$= [0, 0, 0, 1]\cdot\frac{1}{2}[0, 1, 1, -1] = \frac{x_1 + x_2 - |x_1 - x_2|}{2}.$$

It is noteworthy that the last expression has been known since 1953 [3], but the author didn't describe the method of its obtaining (more precisely, it was described later).

The authors of this article propose three approaches to the creation of the basis vectors of linear bases, depending on the operations used for this purpose:

- based on Boolean operations ∨ , &, ⊕, ~, etc .;

- based on the truncated difference operation ∸;

- based on the comparison operation ≥.

When using Boolean operations, the upper and lower "cuts" [14, 18] of variables of different orders are used as operands for creating basis vectors:

– variables of the first order $x^{(i)} = x\&i,\ x_{(i)} = x \vee i$;

– variables of the second order

$$x^{(ij)} = x\&i - x\&j,\ x_{(ij)} = x \vee i - x \vee j$$

and so on.

When using the *truncated difference* operation

$$x_1 \dot{-} x_2 = \begin{cases} x_1 - x_2 & \text{when } x_1 \geq x_2;\\ 0 & \text{when } x_1 < x_2,\end{cases}$$

all logical operations are replaced by some combinations of this operation on variables involving operations of linear space. For example, the upper and lower "cuts" of the first and second orders are replaced by the expressions

$$x^{(i)} = i \dot{-} (i \dot{-} x) = x \dot{-} (x \dot{-} i);$$

$$x_{(i)} = i + (x \dot{-} i) = x + (i \dot{-} x);$$

$$x^{(ij)} = (x \dot{-} j) \dot{-} (x \dot{-} i);$$

$$x_{(ij)} = (i - j) \dot{-} (x - j).$$

Similarly, Boolean operations on two variables are replaced by the following expressions:

$$x_1\&x_2 = x_1^{(x_2)} = x_2 \dot{-} (x_2 \dot{-} x_1) = x_1 \dot{-} (x_1 \dot{-} x_2);$$

$$x_1 \vee x_2 = x_{1(x_2)} = x_2 + (x_1 \dot{-} x_2) = x_1 + (x_2 \dot{-} x_1);$$

$$x_1 \oplus x_2 = x_1 + x_2 - k\{1 \dot{-} [k \dot{-} (x_1 + x_2)]\};$$

$$x_1 \ominus x_2 = x_1 - x_2 + k\{1 \dot{-} [1 \dot{-} (x_2 \dot{-} x_1)]\};$$

and others.

When using the *comparison* operation, the above expressions are reduced to the following form

$$x^{(i)} = \sum_{j=0}^{i-1}(x > j);$$

$$x^{(i)} - x^{(j)} = \sum_{t=j}^{i-j}(x > t);$$

$$x_{(i)} = x + \sum_{j=0}^{i}(j > x) = i + \sum_{j=i+1}^{k-2}(x > j);$$

$$x_{(i)} - x_{(j)} = (i-j) - \sum_{t=j}^{i-1}(x > t).$$

It should be noted that the most interesting result of the studies is the fact of constructing logical structures based on *truncated difference* and *comparison* operations other than Boolean ones. Naturally, there are certain dependencies between the *truncated difference* and *comparison* operations, some of which are given below:

-"*truncated difference – comparison*":

$$x \dot{-} i = \sum_{j=i+1}^{k-1}(x > j);$$

$$i \dot{-} x = \sum_{j=1}^{i}(j > x);$$

$$j \dot{-} (x \dot{-} i) = \sum_{p=1}^{j}[(i+p) > x];$$

$$j \dot{-} (i \dot{-} x) = (j \dot{-} i) + \sum_{p=i-1}^{i \dot{-} j}(x > p);$$

- "*comparison - truncated difference*":

$$x \le i = 1 \dot{-} \{1 \dot{-} [(i+1) \dot{-} x]\};$$

$$x < i = 1 \dot{-} [1 \dot{-} (i \dot{-} x)];$$

$$x \ge i = 1 \dot{-} [1 \dot{-} (x \dot{-} i)];$$

$$x > i = 1 \dot{-} [(i+1) \dot{-} x];$$

$$x_1 > x_2 = 1 \dot{-} [1 \dot{-} (x_1 \dot{-} x_2)];$$

$$x_1 \ge x_2 = 1 \dot{-} \{1 \dot{-} [(x_1 \dot{-} x_2) + 1]\};$$

$$x_1 < x_2 = 1 \dot{-} \{1 \dot{-} [(x_2 \dot{-} x_1) + 1]\}.$$

3. Monotonic Logic Functions

Regarding the close connection between threshold and monotonic functions, we present some known results obtained in linear algebra in a simpler and more obvious way.

A. Construction of sequences of nondecreasing componentes. Suppose we have an arbitrary vector $a = (a_0, a_1, , \dots, a_{m-1}) \in Z^m$. We renumber the components of the vector with k-ary n-bit numbers. The vector a$\in$ Z^m is called monotonically increasing (decreasing), if for a bitwise comparison of k-ary number codes of the components a_i and a_j we have:

$$\forall (i, j \in Z^m) i \ge j \Rightarrow a_i \ge a_j \wedge i \le j \Rightarrow a_i \le a_j. \qquad (2)$$

The bitwise comparison of k-ary number codes of the components enables to single out the sequences of nondecreasing (nonincreasing) components.

The necessity to create these sequences is that for the monotonicity of the vector, condition (2) must be met in each sequence. Hence follows the first factor of simplicity of the analysis results in linear algebra: to verify the vector by monotonicity it is sufficient to establish its monotonicity within each sequence. This reduces the amount of computation and the overall complexity of the verification process.

The sequences of nondecreasing components are constructed as follows. The set G of components (constituents 1) of logic functions of the chosen number of n variables is divided into n groups $G = \{g_0, \dots, g_{n-1}\}$. Each group g_{it} includes the sequences of nondecreasing components with starting numbers t, $i \le t \le n-1$, determined by the following relation

$$k^i - k^{i-1} \le j \le k^{i+1} - k^i - 1,$$

in this case we take $(k^i - k^{i-1})\big|_{i=0} = 0$. The starting numbers are component numbers that should be compared with the values of the other elements in the sequence. Each sequence contains components with numbers the decimal equivalents of which are determined as $t + k^t, t + k^{t+1}, \dots, t + k^{n-1}$. As a result, the entire sequence generally takes on the form

$$g_{it} = \{g_t, g_t + k^t, g_t + k^{t+1}, \dots, g_t + k^{n-1}\}.$$

Thus, it is possible to construct the sequences of nondecreasing components for the given values of k and n. The sequence graphs of nondecreasing components for $k = 2$, $k = 3$ and $n = 3$ are shown in Figure 1. The digits in the circles denote the decimal numbers of the vector components, and the sequence itself includes a certain vertex and the nearest right vertexes connected with it by the lines.

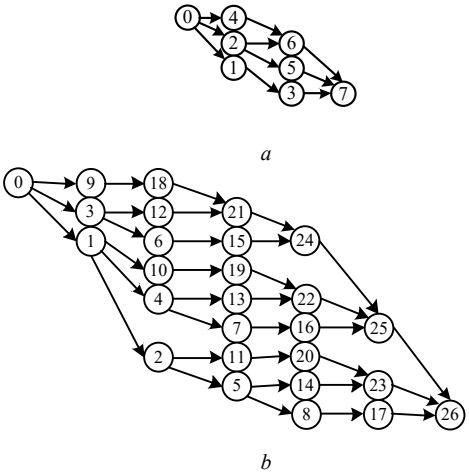


Figure 1: Sequence graphs of the nondecreasing components for $k = 2$, $k = 3$ and $n = 3$.

Example 2. Construct sequences of nondecreasing components for $k = n = 3$. Since $n = 3$, there are 3 groups of sequences with starting numbers 0, 1 and 2 for the initial data under consideration.

For group "0", the value of t_0 varies in the range

$$3^0 - 3^{-1} \le t_0 \le 3^1 - 3^0 - 1,$$

or $\quad 0 \le t_0 \le 1.$

Therefore, this group includes two sequences with the initial elements $t_{00} = 0$ and $t_{01} = 1$. The first sequence g_{00} of the group "0" contains the elements with numbers t_{00}, $t_{00} + 3^0$, $t_{00} + 3^1$, $t_{00} + 3^2$, i.e. 0, 1, 3, 9. The second sequence g_{01} of the group "0" comprises the elements with numbers t_{01}, $t_{01} + 3^0$, $t_{01} + 3^1$, $t_{01} + 3^2$, i.e. the elements with numbers 1, 2, 4, 10.

Similarly, for group 1, the value of t_1 is

$$3^1 - 3^0 \le t_1 \le 3^2 - 3^1 - 1,$$

or

$$2 \le t_1 \le 5.$$

Consequently, this group includes sequences with the initial elements $g_{1t} = 2 ... 5$, and each sequence of the group comprises the elements with numbers t_{1j}, $t_{1j} + 3^1$, $t_{1j} + 3^2$.

Finally, for group 2, the value of t_2 is

$$3^2 - 3^1 \le t_2 \le 3^3 - 3^2 - 1,$$

or

$$6 \le t_2 \le 17,$$

i.e. this group contains sequences with the initial elements $g_{2t} = 6 ... 17$, and each sequence includes the elements with numbers t_{2j}, $t_{2j} + 3^2$. Thus, we obtain the sequence structure of nondecreasing components in the form shown in Figure. 1*b*.

The monotonicity property of vectors from Z^m enables to obtain various representations of an arbitrary vector by means of monotonic vectors. Let's consider such representations in the following three variants:

– the difference of two monotonic vectors that have a value greater than the significance of the represented vector;

– the algebraic sum of monotonic vectors of the same value as the original vector;

– the algebraic sum of monotonic vectors of the value, the smaller value of the represented vector.

B. The problem of representing the logic function by the difference of two monotonic vectors of greater significance is solved by the following algorithm [18]:

Algorithm 1.

1. i=0. $b_0 = a_0$, $c_0 = b_0 - a_0$;

2. $1 \le i \le m - 1$.

$$b_i = \begin{cases} b_{i-1} \text{ when } a_i < a_{i-1} \\ b_{i-1} + a_i - a_{i-1} \text{ when } a_i \ge a_{i-1} \end{cases}, \quad c_i = b_i - a_i;$$

3. $i = m$, the end.

Here i and i–1 are neighboring indices of the elements in the sequence of nondecreasing components, $m = k^n$. The validity of the algorithm results from the following elementary arguments.

1. Let $b_0 = a_0$, then $c_0 = b_0 - a_0 = 0$.

2. Let the pair of components a_δ and a_γ satisfies the condition of monotonicity, i.e. $a_\delta \le a_\gamma$. In this case

$$a_\delta - a_\gamma \ge 0,$$

and, consequently, from the identity

$$b_\gamma = b_\delta + a_\gamma - a_\delta,$$

it follows that

$$b_\gamma \ge b_\delta,$$

i.e. b_γ also meets the monotonicity condition.

3. If the pair of components a_γ and a_δ doesn't fulfill the condition of monotonicity, i.e. $a_\delta \ge a_{\gamma\gamma}$, then, taking $b_\gamma = b_\delta$ we remain the monotonicity condition for the components of the vector b again.

Example 3. Construct a representation of the vector a = {0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0} as a difference of two monotonic vectors for $k = 2$, $n = 4$, $m = k^n = 16$.

The sequence of nondecreasing components for this case has the following form:

Table 1

g	Sequences	Values of components
0	0, 1, 2, 4, 8	0, 1, 1, 1, 1
1	1, 3, 5, 9,	1, 0, 0, 0
2	2, 6, 10	1, 0, 1
	3, 7, 11	0, 1, 1
3	4, 12	1, 1
	5, 13	0, 1
	6, 14	0, 1
	7, 15	1, 0

We construct the components of the vectors b and c:

$$b_0 = 0 \qquad c_0 = 0$$
$$b_1 = b_1 + a_1 - a_0 = 1 \qquad c_1 = 0$$
$$b_2 = b_1 + a_2 - a_1 = 1 \qquad c_2 = 0$$
$$b_3 = b_1 = 1 \qquad c_3 = 1$$
$$b_4 = b_2 + a_4 - a_2 = 1 \qquad c_4 = 0$$
$$b_5 = b_1 = 1 \qquad c_5 = 0$$
$$b_6 = b_2 = 1 \qquad c_6 = 1$$
$$b_7 = b_3 + a_7 - a_3 = 2 \qquad c_7 = 1$$
$$b_8 = b_4 + a_8 - a_4 = 1 \qquad c_8 = 0$$
$$b_9 = b_5 + a_9 - a_5 = 1 \qquad c_9 = 1$$

$$b_{10} = b_6 + a_{10} - a_6 = 2 \qquad c_{10} = 1$$
$$b_{11} = b_7 + a_{11} - a_7 = 2 \qquad c_{11} = 1$$
$$b_{12} = b_8 + a_{12} - a_8 = 1 \qquad c_{12} = 0$$
$$b_{13} = b_9 + a_{13} - a_9 = 2 \qquad c_{13} = 1$$
$$b_{14} = b_6 + a_{14} - a_6 = 1 \qquad c_{14} = 1$$
$$b_{15} = b_7 = 2 \qquad c_{15} = 2$$

The described algorithm demonstrates one more fact, which confirms the simplicity of analysis in linear algebra.

As it can be seen from the example, when using the described algorithm in the general case, the value of the original vector does not coincide with the significance of the resolution vectors.

C. The expansion of the arbitrary logic function of n arguments into the algebraic sum of monotonic logic functions of the same value of the following form

$$f(\tilde{x}^{(n)}) = (-1)^i \sum_{i=0}^{p} \varphi_i(\tilde{x}^{(n)}),$$

where $p \le kn$, $\varphi_i(\tilde{x}^{(n)})$ – monotonic expansion functions with the property $\varphi_1 \supset \varphi_2 \supset \ldots$, are made in accordance with the following algorithm [18].

Algorithm 2.

1. We choose a minimal summation of the positive summands of the arithmetic-logical representation of the logic function $f(\tilde{x}^{(n)})$ covering (in the logical sense) all other positive summands, and combine it by the operation $\vee$. Thus, we form the first expansion function $\varphi_1(\tilde{x}^{(n)})$. To remain the equality, all possible logical products of summands from $\varphi_1(\tilde{x}^{(n)})$ with the signs defined as $(-1)^j$, where $1 < j < s$, s is a number of summands in $\varphi_1(\tilde{x}^{(n)})$, are added to $\varphi_1(\tilde{x}^{(n)})$. Then we reduce similar terms and represent the initial logic function $f(\tilde{x}^{(n)})$ in the following form $f(\tilde{x}^{(n)}) = \varphi_1(\tilde{x}^{(n)}) - f_1(\tilde{x}^{(n)})$, where $f_1(\tilde{x}^{(n)})$ is a remainder of the initial function after reduction of similar terms.

2. We repeat clause 1 for $f_1(\tilde{x}^{(n)})$. As a result, the initial logic function is represented in the following form

$$f(\tilde{x}^{(n)}) = \varphi_1(\tilde{x}^{(n)}) - \varphi_2(\tilde{x}^{(n)}) + f_2(\tilde{x}^{(n)}).$$

3. We repeat clause 2 until the remainder of the initial logic function becomes zero.

Since the number of arguments of the logic function is n, and the violation of monotonicity is possible in each of the k values of each argument, then, the maximum number of the expansion functions doesn't exceed $2n(k–1)$.

The process convergence follows from the fact that each successive resolution vector eliminates some violation of monotonicity in the original vector and doesn't introduce new monotonicity violations, since it is monotonic itself.

The problem of representing the arbitrary logic function by the algebraic sum of monotonic functions of the same value can be solved analytically. As is known [13], the minimal disjunctive normal form (DNF) of the monotonic logic function doesn't contain inversions over variables. Consequently, the given logic function must be reduced to the representation in the form of the algebraic sum of such functions. To do this, we must perform the following actions:

– conversion of the Boolean expression of the logic function into the linear one using the following identities

$$x_1 \vee x_2 = x_1 + x_2 - x_1 x_2;$$

$$\bar{x} = 1 - x;$$

– inverse transformation to the given (that doesn't contain inversions over variables) form.

Example 4. Obtain the two-valued logic function mapping of three arguments

$$f(\tilde{x}^{(3)}) = (0,1,1,0,1,0,0,1),$$

into the algebraic sum of monotonic logic functions.

Solution. The logic function is represented by a vector of values. Its linear representation has the following form

$$f(\tilde{x}^{(3)}) = x_1 + x_2 + x_3 - 2x_1x_2 - 2x_1x_3 - 2x_2x_3 + 4x_1x_2x_3.$$

Now we obtain the mapping of this function into the algebraic sum of monotonic functions. The first expansion function is formed from the first three summands, covering in aggregate all the remaining positive terms of sum:

$$f(\tilde{x}^{(3)}) = (x_1 \vee x_2 \vee x_3 + x_1x_2 + x_1x_3 + x_2x_3 - x_1x_2x_3) -$$
$$-2x_1x_2 - 2x_1x_3 - 2x_2x_3 - 4x_1x_2x_3 =$$
$$= \varphi_1(\tilde{x}^3) - x_1x_2 - x_1x_3 - x_2x_3 + 3x_1x_2x_3.$$

The subsequent expansion functions are obtained in the same way:

$$f(\tilde{x}^{(3)}) = \varphi_1(\tilde{x}^{(3)}) - (x_1x_2 \vee x_1x_3 \vee x_2x_3 + x_1x_2x_3 +$$
$$+x_1x_2x_3 + x_1x_2x_3 + x_1x_2x_3 - x_1x_2x_3) + 3x_1x_2x_3 =$$
$$= \varphi_1(\tilde{x}^{(3)}) - \varphi_2(\tilde{x}^{(3)}) + \varphi_3(\tilde{x}^{(3)}).$$

To obtain the threshold representation of the initial function, it suffices to transform the monotonic functions of the resulting expansion into the threshold form and to perform their algebraic addition.

Each conjunctive term of any logical expansion function is transformed into the threshold form in accordance with the following identical equation

$$x_i x_j \ldots x_s = P[(x_i + x_j + \cdots + x_s) > t - 1],$$

where t - a number of arguments in the term, P - the predicate symbol. Then all the terms are reduced to the common value of the right-hand side with the introduction of the corresponding

coefficients for the variables, then they are added together and the threshold value of the received sum is determined.

Let's explain the transformation of the monotonic logic function into the threshold form with the function from the previous example.

With the transformation of $\varphi_1(\tilde{x}^{(3)})$ everything is simple:

$$\varphi_1(\tilde{x}^{(3)}) = x_1 \vee x_2 \vee x_3 = P_1[(x_1 + x_2 + x_3) > 0];$$

The transformation of $\varphi_2(\tilde{x}^{(3)})$ looks somewhat more complicated:

$$\varphi_2(\tilde{x}^{(3)}) = x_1x_2 \vee x_1x_3 \vee x_2x_3 =$$
$$= P_2[(x_1 + x_2) > 1 + (x_1 + x_3) > 1 + (x_2 + x_3) > 1] =$$
$$P_2[(2x_1 + 2x_2 + 2x_3) > 2] = P_2[(x_1 + x_2 + x_3) > 1];$$

The transformation of $\varphi_3(\tilde{x}^{(3)})$ is produced just as easily as $\varphi_1(\tilde{x}^{(3)})$:

$$\varphi_3(\tilde{x}^{(3)}) = x_1x_2x_3 = P_3[(x_1 + x_2 + x_3) > 2].$$

Thus, to realize the considered logic function, three threshold logic elements are required. The simplicity of the transformations in linear algebra becomes obvious.

D. Representation of the arbitrary vector by the vectors of lower significance. This problem has many solutions. The variant considered below assumes the solution of this problem in two stages:

- obtaining a representation of the arbitrary vector from Z^m by the vectors of lower significance;

- transformation of the obtained representation into the representation by means of monotonic vectors.

Let us first consider the solution of the first stage of the problem. The representation of the arbitrary vector by the lower-valued vectors can be obtained by weighting (each resolution vector is included into the final representation with some weight coefficient) or unitary (each resolution vector is included into the final representation with a unitary weight coefficient) coding.

1. It is known from the theory of numbers [33] that any number a can be uniquely represented in the following form:

$$\sum_1^s a_{s-i}k^{s-i}, \qquad (3)$$

where $s = [\log_k a]$ – the closest integer to $\log_k a$, a_{s-i} – the values of the expansion coefficients of the number a in the k-valued number system. Due to the uniqueness, this relation determines the isomorphism between any number a and the described representation of this number.

The quantity a_{s-i} can be determined from the following relation:

$$a_{s-i} = \frac{a\,(mod\ k^{s-i+1}) - a\,(mod\ k^{s-i})}{k^{s-i}}.$$

This operation is linear

$$a + b = \sum_{i=1}^{s} a_{s-i}\,k^{s-i} + \sum_{i=1}^{s} b_{s-i}\,k^{s-i} =$$
$$= \sum_{i=1}^{s}(a_{s-i} + b_{s-i})\,k^{s-i};$$
$$\lambda a = \sum_{i=1}^{s} \lambda a_{s-i}\,k^{s-i} = \lambda \sum_{i=1}^{s} a_{s-i}\,k^{s-i},$$

thus, it is applicable to the vectors of the linear space. If now we associate the vector a to the number a in the linear space Z^m, and the weighted sums of the vectors a_{s-i} of the spaces K_1^m and K_2^m to the weighted sums $\sum_{i=1}^{s_1} a_{s_1-i}\,k^{s_1-i}$ and $\sum_{i=1}^{s_2} a_{s_2-i}\,k^{s_2-i}$ correspondingly, then the expression (3) is equivalent to the representation of the vector $a \in Z^m$ by the weighted sum of the vectors from K_1^m or K_2^m respectively. Since in both spaces the vector a is uniquely represented, both representations are isomorphic. Let's call such a representation by coding, and the corresponding linear operator - by *the encoding operator.*

Let the vector $a \in Z^m$ in the canonical basis be represented as

$$a = a_1e_1 + a_2e_2 + \cdots + a_me_m.$$

Using expression (3), we expand each coordinate of the vector a with respect to the bases *k1* and *k2*. As a result, we obtain its representation in the following form

$$a_i = \sum_{i=1}^{s_1} a_{i,s_1-i}\,k_1^{s_1-i},$$

or

$$a_j = \sum_{j=1}^{s_2} a_{j,s_2-j}\,k_2^{s_2-j},$$

where $a_{i,s-i}$ – the expansion components of the number a_i (i – e vector components a_i of the vector representation a).

Then the vector representation a in each of the spaces can be described in the following form

$$a = \left(a_{1,s-1}k^{s-1} + a_{1,s_1-2}k^{s-2} + \cdots + a_{1,0}k^0\right) +$$
$$+\left(a_{2,s-1}k^{s-1} + a_{2,s_1-2}k^{s-2} + \cdots + a_{2,0}k^0\right) + \cdots$$
$$+ \left(a_{m,s-1}k^{s-1} + a_{m,s_1-2}k^{s-2} + +a_{m,0}k^0\right),$$

or

$$a = \left(a_{1,s-1} + a_{2,s-1} + \cdots + a_{m,s-1}\right)k^{s-1} +$$
$$+\left(a_{1,s-2} + a_{2,s-2} + \cdots + a_{m,s-2}\right)k^{s-2} + \cdots$$
$$+\left(a_{1,0} + a_{2,0} + \cdots + a_{m,0}\right)k^0.$$

Thus, the arbitrary vector a can be written in the following form

$$a = \sum_{j=1}^{s}\left(\sum_{i=1}^{m} a_{i,s-j}\right)k^{s-j},$$

i.e. a is represented by the weighted sum s of the vectors from K_1^m or K_2^m

Example 5. Let $k_1 = 10$, $n = 1$. The problem is to represent the vector $a = \{0, 1, 3, 9, 7, 6, 9, 4, 3, 8\}$ by its k_2-weighted and k_2-unitary sums of the vectors when $k_2 = 3$.

Solution. Since $\lceil log_{k_2} k_1 \rceil = \lceil log_3 10 \rceil = 2$, then the vector a can be represented by the weighted sum of three vectors a_2, a_1 and a_0. The coordinates of the highest weight vector a_2 are determined as

$$a_2 = \frac{a\ mod 3^3 - a\ mod 3^2}{3^2} =$$

$$= \frac{(0139769348) - (0130760438)}{9} = (0001001000).$$

The components of the subsequent vectors a_1 and a_0 are determined similarly:

$$a_1 = \frac{a\ mod 3^2 - a\ mod 3^1}{3^1} =$$

$$= \frac{(0130760348) - (0100100102)}{3} = (0010220112);$$

$$a_0 = \frac{a\ mod 3^1 - a\ mod 3^0}{3^0} =$$

$$= \frac{(0100100102) - (0000000000)}{1} = (0100100102).$$

Thus,

$$a = \sum_0^3 a_i 3^i = (0001001000)3^2 + (0010220112)3^1 + (0100100102)3^0.$$

2. We now represent the arbitrary number a as follows:

$$a = \sum_{i=0}^{s-1}(a - \sum_{i=0}^{s-1} a_i)\ \&(k_2 - 1), \qquad (4)$$

where s - the nearest larger integer

$$s = \left\lceil \frac{k_1 - 1}{k_2 - 1} \right\rceil.$$

For the reasons stated above, this representation is also an isomorphism.

If we associate each coordinate of the vector a with its representation in form (4), we obtain:

$$a = \left[a_1\&(k_2 - 1) + \left(a_1 - a_{11}\&(k_2 - 1)\right) + \cdots + a_1 - \sum_{i+1}^{s-1} a_{1i}\right]\&(k_2 - 1) +$$

$$+ \left[a_2\&(k_2 - 1) + \left(a_2 - a_{21}\&(k_2 - 1)\right) + \cdots + a_2 - \sum_{i+1}^{s-1} a_{2i}\right]\&(k_2 - 1) + \cdots$$

$$+ \left[a_m\&(k_2 - 1) + \left(a_m - a_{m1}\&(k_2 - 1)\right) + \cdots + a_m - \sum_{i+1}^{s-1} a_{mi}\right]\&(k_2 - 1).$$

Thus,

$$a = \sum_0^{s-1} a_i,$$

where $a_i = (a - \sum_{i+1}^{s-1} a_i)\&(k_2 - 1)$, i.e. the arbitrary vector $a \in Z^m$ is represented by the k_2-unitary sum of vectors from Z^m.

Example 6. Present the original vector from the previous example using unitary coding.

Solution. Since under the same initial conditions $s = \left\lceil \frac{10-1}{3-1} \right\rceil = 4$, the vector a can be represented by the unitary sum of five vectors a_4, a_3, a_2, a_1 and a_0, where

$a_4 = a\&(3 - 1) = (0139769438)\&(3 - 1) = (0122222222)$;

$a_3 = (a - a_4)\&(3 - 1) = (0012222212)$;

$a_2 = (a - a_4 - a_3)\&(3 - 1) = (0002222002)$;

$a_1 = (a - a_4 - a_3 - a_2)\&(3 - 1) = [0002102002]$;

$$a_0 = (a - a_4 - a_3 - a_2 - a_1)\&(3 - 1) =$$

$$= [0001001000].$$

Thus,

$$a = \sum_0^3 a_i = (0122222222) + (0012222212) +$$

$$+(0002222002) + (0002102002) +$$

$$+0001001000.$$

The solution of the second stage of the problem, i.e. the transformation of the obtained representation into the representation by means of monotonic lower-valued vectors, can be obtained on the base of algorithm 2.

Thus, the body of the linear space theory determines easy-to-use means of representing arbitrary vectors of any given value by monotonic vectors of greater, similar or lesser significance, and also converting monotonic functions to the threshold form.

4. Synthesis of Logic and Threshold Elements

In linear algebra, the logic function is represented by a weighted algebraic sum of terms. As a whole the variables, terms and function take values from the set of numbers 0, 1, ..., k - 1. The algebraic sum is realized by operations of the linear space, and individual terms - depending on the choice made - by other operations chosen for technical, technological or operational reasons. In this article, *"Truncated Difference"* and *"Comparison"* are used as such operations. In the future, the developers of linear LSIs can be motivated to choose other operations. Thus, the process and the possibilities of logic synthesis will be further demonstrated on the bases based on the *truncated difference* and *comparison* operations.

To solve the problems of logic synthesis in linear algebra, first, it is necessary to solve the problem of the formation of bases.

In Boolean algebra, 17 previously defined functionally complete systems of logic functions of two arguments are known [34]. Selecting any of them, you can get its linear analog and

perform a logic synthesis of the circuit that implements the given logic function according to the established algorithm. Therefore, it is possible to construct bases as a mapping of Boolean functionally complete systems to a linear space. For *truncated difference* and *comparison* operations, the bases can be formed directly.

To form the bases, we need:

- a set of operations;

- a set of generating functions.

The following operation are considered as the basic ones:

- logical operations in their traditional understanding;

- operation *"truncated difference";*

- operation *"comparison".*

Constants 1, logical variables x_1, x_2, …, x_n and their totals subjected to the effect of basic operations are used as generators for the formation of bases.

The following methods of logic synthesis of digital structures are considered:

- direct synthesis in a given basis (logic synthesis in the linear space);

- synthesis with preliminary transformation of the value of the given function (for example, a multivalued two-digit implementation);

- synthesis with preliminary expansion of the given function into the algebraic sum of monotonic functions (threshold synthesis).

The first method is the most obvious and consists in the representation of the realized logic function in the chosen basis.

The second method presupposes a preliminary representation of the realized function in the form of an algebraic sum of logic functions larger, smaller or equal to the original value. The methods of this transformation have been described in the previous section.

Finally, the third method consists of the preliminary representation of the realized function by the algebraic sum of monotonic logic functions. In this case, each monotonic function is realized by a single threshold element.

Note that the synthesis of digital structures with increasing of their complexity, as usual, requires the involvement of methods of system engineering.

A. Basic operations of logic synthesis. The set of operations most often used to represent logic functions will be considered as basic operations of logic synthesis. In Boolean algebra, such operations can be considered as operations of the basic functional system AND (&), OR (∨), ($\bar{}$), and also the operations "modulo 2 sum" (⊕), "logic equivalence" (~) "dual stroke" (↓), "Sheffer stroke" (|), "prohibition with respect to x_1 (x_2)" $(x_1\bar{x}_2)$, "implication from x_1 to x_2)" $(x_1 \vee \bar{x}_2)$, majority operation $x_1 \# x_2 \# x_3$.

By *basic operations of linear algebra* we mean arithmetic operations (operations of the linear space), as well as operations used to form bases. We classify the current realizations of logical operations: the difference module, *the truncated difference*, the arithmetic sum and the difference, the multiplication by a constant coefficient, the operations of changing the sign to them. Perhaps there are other arithmetic operations suitable for the use in linear logic synthesis, but they are not considered in this article.

Since our task is logic synthesis in linear algebra, we start with the logic synthesis of basic operations of linear algebra.

Algebraic sum. The operation is realized by the assembly connection of conductors, over which the currents flow. The conditional graphic representation of the operation of the algebraic sum is shown in Figure 2.

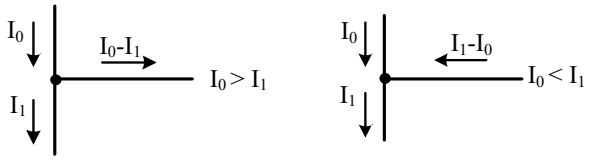


Figure 2: Conditional graphical representation of the algebraic addition of currents.

The arithmetic sum differs from the algebraic one in that in the latter case the signals of only one sign (i.e. only inflowing or only flowing out) are fed to the input of the element. From the output, the sum of the signals is removed. The positive direction of the input currents is the current direction to the connection point Σ_1, i.e. the positive current flowing into the connection point. The negative direction of the input currents is the direction of the current from the connection point of the conductors Σ_1, i.e. negative - the current flowing from the connection point. For the output (resulting) current, the positive direction is the direction from the connection point Σ_1, and the negative direction - to the connection point Σ_1,

The number of positive and negative inputs is determined by the function being implemented. Since the logical variables take values from the positive semi-axis of the numerical axis, when subtracting, it is necessary to meet the condition that the total sum of the negative summands (current quanta) doesn't exceed the total sum of the positive terms of the sum. Otherwise, instead of the algebraic sum, the *truncated difference* operation is performed, in which the subtraction from zero is a logical component of this operation and which carries the meaning of the *comparison* operation with zero (Figure 3)

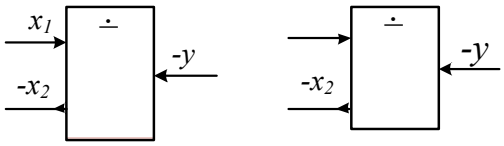
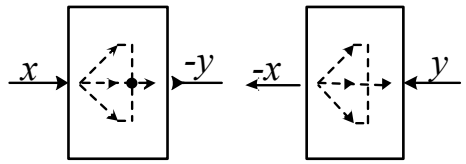


Figure 3: Conditional graphic representation o*f the truncated difference* operation.

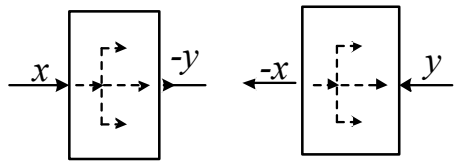
The operation of multiplying (dividing) by a constant coefficient consists in multiplying the input signal by several outputs and then combining the outputs of the multiplied signal (when multiplying) or outputting some part of the input signal (when dividing), as shown in Figure 4.

The comparison operation is used to determine the relationship between the compared variables. It can be performed in two forms: in the relative form, i.e. in the form of determining the difference in the values of the variables, or in the absolute form, i.e. in the form of the "more-less" comparison itself. The conditional graphic

symbols of the comparison elements in these forms are shown in Figure 5.

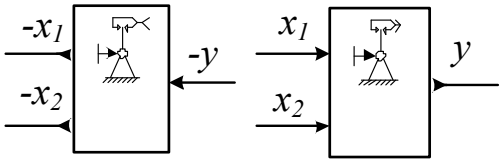


a

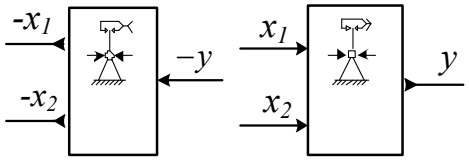


b

Figure 4: Conditional graphic representation of the operation of multiplication *(a)* and division *(b)*.



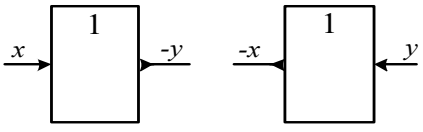
a



b

Figure 5: Conditional graphic representation of the *comparison* operation: relative *(a)* and absolute *(b)*.

The operation of *changing the sign* of terms allows changing the sign of the term, thereby transforming it from the summand to the subtrahend and vice versa. It is performed by the current inverter, the conditional graphic representation of which is shown in Figure 6.



a *b*

Figure 6: Conditional graphic representation of the operation of changing the sign.

The operation of changing the sign can functionally be combined with the operations of multiplication and division by the constant coefficient.

B. Formation of bases. The bases being formed are bases of the linear space, and according to the form they are sets of variables and their various combinations (terms) combined by the selected operations. We can approach to the formation of bases from different sides. For the two-valued synthesis, as it was shown above, it is possible, for example, to construct bases equivalent to 17 functionally complete systems (logical-arithmetic), and then create their analogs in linear algebra.

To form the multivalued bases, we can use the cuts of multivalued variables (10 combinations of such cuts are proposed in [18]), and then construct their analogs in linear algebra. Such a solution is also applicable for the two-valued bases. The problem of forming the bases is solved uniformly for the cuts of any type, so we demonstrate its solution on the basis of the upper cuts of the first level, which are analogs of the functionally complete system AND, NOT (OR, NOT).

The bases of this type for $k = 2$, $n = 2$ and $n = 3$, and for $k = 3$, $n = 2$ are given below. For each basis, the basic and inverse matrices are presented. Two-digit operations & and ∨ and their multivalued analogs $min(x_1x_2)$ and $max(x_1x_2)$ are used as generating operations. Therefore, the bases of this type are called logical-arithmetic.

The two-valued logical-arithmetic bases of two variables have the form:

- conjunctive (AND, NOT)

$$\&(\mathrm{B1},\mathrm{B1})^{(2)} = \begin{bmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \\ x_1^{(1)}x_2^{(1)} \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1\&x_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$

$$\left(\&(\mathrm{B1},\mathrm{B1})^{(2)}\right)^{-1} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & 1 & 1 \end{bmatrix};$$

- disjunctive (OR, NOT)

$$\vee\,(\mathrm{B1},\mathrm{B1})^{(2)} = \begin{bmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \\ x_1^{(1)} \vee x_2^{(1)} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix};$$

$$(\vee\,(\mathrm{B1},\mathrm{B1})^{2})^{-1} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & -1 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 1 & 1 & -1 \end{bmatrix}.$$

The two-valued logical-arithmetic bases of three variables have the similar form:

$$\&(\mathrm{B1},\mathrm{B1},B1)^{(2)} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_1x_2 \\ x_1x_3 \\ x_2x_3 \\ x_1x_2x_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

$$\left(\&(\mathrm{B1},\mathrm{B1},B1)^{(2)}\right)^{-1} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

$$\vee(\mathrm{B1},\mathrm{B1},B1)^{(2)}=\begin{bmatrix}1\\x_1\\x_2\\x_3\\x_1\vee x_2\\x_1\vee x_3\\x_2\vee x_3\\x_1\vee x_2\vee x_3\end{bmatrix}=\begin{bmatrix}1&1&1&1&1&1&1&1\\0&1&0&1&0&1&0&1\\0&0&1&1&0&0&1&1\\0&0&0&0&1&1&1&1\\0&1&1&1&0&1&1&1\\0&1&0&1&1&1&1&1\\0&0&1&1&1&1&1&1\\0&1&1&1&1&1&1&1\end{bmatrix};$$

$$\left(\vee(\mathrm{B1},\mathrm{B1},B1)^{(2)}\right)^{-1}=\begin{bmatrix}1&0&0&0&0&0&0&-1\\0&0&0&0&0&0&-1&1\\0&0&0&0&0&-1&0&1\\0&0&0&-1&0&1&1&-1\\0&0&0&0&-1&0&0&1\\0&0&-1&0&1&0&1&-1\\0&-1&0&0&1&1&0&-1\\0&1&1&1&-1&-1&-1&1\end{bmatrix}.$$

The three-valued logical-arithmetic analogs of these bases look like this:

- conjunctive

$$min(B1,B1)^{(3)}=\begin{bmatrix}1\\x_1^{(1)}\\x_1^{(2)}\\x_2^{(1)}\\min\left(x_1^{(1)},x_2^{(1)}\right)\\min\left(x_1^{(2)},2x_2^{(1)}\right)\\x_2^{(2)}\\min\left(2x_1^{(1)},x_2^{(2)}\right)\\min\left(x_1^{(2)},x_1^{(2)}\right)\end{bmatrix}=\begin{bmatrix}1&1&1&1&1&1&1&1&1\\0&1&1&0&1&1&0&1&1\\0&1&2&0&1&2&0&1&2\\0&0&0&1&1&1&1&1&1\\0&0&0&0&1&1&0&1&1\\0&0&0&0&1&2&0&1&2\\0&0&0&1&1&1&2&2&2\\0&0&0&0&1&1&0&2&2\\0&0&0&0&1&1&0&1&2\end{bmatrix};$$

$$\left(min(B1,B1)^{(3)}\right)^{-1}=\begin{bmatrix}1&-1&0&-1&1&0&0&0&0\\0&0&0&2&-2&0&-1&1&0\\0&0&0&-1&1&0&1&-1&0\\0&2&-1&0&-2&1&0&0&0\\0&0&0&0&2&-1&0&-1&1\\0&0&0&0&0&0&0&1&-1\\0&-1&1&0&1&-1&0&0&0\\0&0&0&0&0&1&0&0&-1\\0&0&0&0&-1&0&0&0&1\end{bmatrix};$$

- disjunctive

$$max(B1,B1)^{(3)}=\begin{bmatrix}1\\x_2^{(1)}\\x_2^{(2)}\\x_1^{(1)}\\max\left(x_2^{(1)},x_1^{(1)}\right)\\max\left(x_2^{(2)},2x_1^{(1)}\right)\\x_1^{(2)}\\max\left(2x_2^{(1)},x_1^{(2)}\right)\\max\left(x_2^{(2)},x_1^{(2)}\right)\end{bmatrix}=\begin{bmatrix}1&1&1&1&1&1&1&1&1\\0&0&0&1&1&1&1&1&1\\0&0&0&1&1&1&2&2&2\\0&1&1&0&1&1&0&1&1\\0&2&2&1&2&2&2&2&2\\0&0&0&0&1&1&0&2&2\\0&1&2&0&1&2&0&1&2\\0&1&2&2&2&2&2&2&2\\0&1&2&1&1&2&2&2&2\end{bmatrix};$$

$$\left(max(B1,B1)^{(3)}\right)^{-1}=\begin{bmatrix}1&0&0&0&-1&0&0&0&0\\0&0&0&0&2&0&0&-1&0\\0&-1&0&0&-1&0&0&1&0\\0&0&0&0&2&-1&0&0&0\\0&0&0&0&-2&1&0&1&-1\\0&2&-1&0&0&0&0&-1&1\\0&0&0&-1&-1&1&0&0&0\\0&0&0&2&0&-1&-1&0&1\\0&-1&1&-1&1&0&1&0&-1\end{bmatrix}.$$

To obtain linear analogs of the above logical-arithmetic bases, we can use the relations that establish the connection between logical operations, on the one hand, and with the operations used to form the bases of linear space, on the other hand. For the bases of the considered type these relations are as follows:

- logical operations - *truncated difference* operation

$$x^{(i)}=i\dot{-}(i\dot{-}x)=x\dot{-}(x\dot{-}i);$$

$$x^{(i)}-x^{(j)}=x^{(ij)}=(x\dot{-}j)\dot{-}(x\dot{-}i);$$

$$min(x_1,x_2)=x_1\dot{-}(x_1\dot{-}x_2)=x_2\dot{-}(x_2\dot{-}x_1);$$

$$max(x_1,x_2)=x_1+(x_2\dot{-}x_1)=x_2+(x_1\dot{-}x_2).$$

- logical operations - *comparison* operation

$$x^{(i)}=\sum_{j=0}^{i-1}(x>j);$$

$$x^{(i)}-x^{(j)}=x^{(ij)}=\textstyle\sum_{t=1}^{i-j}P_t(x>t);$$

$$min(x_1,x_2)=P\{[(x_1>0)+(x_2>0)]>1\}+$$
$$+P\{[(x_1>1)+(x_2>1)]>1\};$$

$$max(x_1,x_2)=P\{[(x_1>0)+(x_2>0)]>0\}+$$
$$+P\{[(x_1>1)+(x_2>1)]>0\}.$$

Using the above relations, we can easily obtain linear analogs of the above bases:

- conjunctive $k = 2$, $n = 2$ and $n = 3$

$$\&(\mathrm{B1},\mathrm{B1})^{(2)}=\begin{bmatrix}1\\x_1^{(1)}\\x_2^{(1)}\\x_1^{(1)}x_2^{(1)}\end{bmatrix}=\begin{bmatrix}1\\x_1\\x_2\\(x_1+x_2)\dot{-}1\end{bmatrix}=\begin{bmatrix}1&1&1&1\\0&1&0&1\\0&0&1&1\\0&0&0&1\end{bmatrix};$$

$$\left(\&(\mathrm{B1},\mathrm{B1})^{(2)}\right)^{-1}=\begin{bmatrix}1&-1&-1&1\\0&1&0&-1\\0&0&1&-1\\0&1&1&1\end{bmatrix};$$

$$\&(\mathrm{B1},\mathrm{B1},B1)^{(2)}=\begin{bmatrix}1\\x_1\\x_2\\x_3\\x_1x_2\\x_1x_3\\x_2x_3\\x_1x_2x_3\end{bmatrix}=\begin{bmatrix}1\\x_1\\x_2\\x_3\\(x_1+x_2)\dot{-}1\\(x_1+x_3)\dot{-}1\\(x_2+x_3)\dot{-}1\\(x_1+x_2+x_3)\dot{-}2\end{bmatrix}=$$

$$=\begin{bmatrix}1&1&1&1&1&1&1&1\\0&1&0&1&0&1&0&1\\0&0&1&1&0&0&1&1\\0&0&0&0&1&1&1&1\\0&0&0&1&0&0&0&1\\0&0&0&0&0&1&0&1\\0&0&0&0&0&0&1&1\\0&0&0&0&0&0&0&1\end{bmatrix};$$

$$\left(\&(\mathrm{B1},\mathrm{B1},B1)^{(2)}\right)^{-1} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

- disjunctive $k = 2$, $n = 2$ and 3

$$\vee(\mathrm{B1},\mathrm{B1})^2 = \begin{bmatrix} 1 \\ x_1^{(1)} \\ x_2^{(1)} \\ x_1^{(1)} \vee x_2^{(1)} \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1 + (x_2 \div x_1) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix},$$

$$(\vee(\mathrm{B1},\mathrm{B1})^2)^{-1} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & -1 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 1 & 1 & -1 \end{bmatrix},$$

$$\vee(\mathrm{B1},\mathrm{B1},B1)^{(2)} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_1 \vee x_2 \\ x_1 \vee x_3 \\ x_2 \vee x_3 \\ x_1 \vee x_2 \vee x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ 1 \div [1 \div (x_1 + x_2)] \\ 1 \div [1 \div (x_1 + x_3)] \\ 1 \div [1 \div (x_2 + x_3)] \\ 1 \div [1 \div (x_1 + x_2 + x_3)] \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

$$\left(\vee(\mathrm{B1},\mathrm{B1},B1)^{(2)}\right)^{-1} =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & 0 & 1 & -1 \\ 0 & -1 & 0 & 0 & 1 & 1 & 0 & -1 \\ 0 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{bmatrix}.$$

Similarly, we obtain linear analogs of multivalued bases:

- conjunctive, $k = 3$, $n = 2$

$$min(B1,B1)^{(3)} = \begin{bmatrix} 1 \\ 1 \div (1 \div x_1) \\ x_1 \\ 1 \div (1 \div x_2) \\ [1 - (1 \div x_1)] - (1 \div x_2) \\ x_1 \div 2(1 \div x_2) \\ x_2 \\ 2[1 - (1 \div x_1)] - (2 \div x_2) \\ x_1 \div (x_1 \div x_2) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 2 \end{bmatrix};$$

$$\left(min(B1,B1)^{(3)}\right)^{-1} =$$

$$= \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 & -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix};$$

- disjunctive

$$max(B1,B1)^{(3)} = \begin{bmatrix} 1 \\ 1 \div (1 \div x_1) \\ x_1 \\ 1 \div (1 \div x_2) \\ 1 \div (x_1 + x_2) \\ x_2 + \{2[1 \div (1 \div x_1)] \div x_2\} \\ x_2 \\ x_1 + \{2[1 \div (1 \div x_2)] \div x_1\} \\ x_1 + (x_2 \div x_1) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 2 & 2 & 1 & 2 & 2 & 2 & 2 & 2 \\ 0 & 1 & 2 & 1 & 1 & 2 & 2 & 2 & 2 \end{bmatrix};$$

$$\left(max(B1,B1)^{(3)}\right)^{-1} =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & -2 & 2 & 0 & 1 & -2 \\ 0 & 0 & 0 & 2 & 0 & -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & -1 & 1 & 0 & 1 & -1 \\ 0 & 2 & -1 & 0 & 0 & -1 & 0 & -1 & 2 \\ 0 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

C. Logic synthesis of basic functions. The resulted conditional graphic designations of operations of the linear space can be used for graphic representation of basic Boolean logic operations in linear algebra and construct functional schemes of logic elements on their base. For this it is sufficient:

- to choose the basis from the listed above;

- to use the previously described technique of logic synthesis in linear algebra.

We choose a conjunctive version of the chosen basis as an object of logic synthesis in which we perform a logic synthesis of the two-valued and three-valued functions of circular shift

$$(x_1 \& x_2) \oplus 1 = [1\ 1\ 1\ 0],$$

and

$$min(x_1, x_2) \oplus 1 = [1\ 1\ 1\ 1\ 2\ 2\ 1\ 2\ 0],$$

and

$$min(x_1, x_2) \ominus 1 = [2\ 2\ 2\ 2\ 0\ 0\ 2\ 0\ 1].$$

In accordance with the above, the first action is to obtain the expansion vector of the value vector of the function in terms of the basis. Multiplying the vectors written above by the columns of the inverse matrix of the bases $\&(\mathrm{B1}, \mathrm{B1})^{(2)}$ and $min(B1, B1)^{(3)}$ results in the following:

$$w[(x_1 \& x_2) \oplus 1] = [0\ 1\ 1\ -2];$$

$$w[min(x_1, x_2) \oplus 1] \Rightarrow [1\ 0\ 0\ 0\ 3\ 0\ 0\ 0\ -2];$$

$$w[min(x_1, x_2) \ominus 1] \Rightarrow [2\ 0\ 0\ 0\ -3\ 0\ 0\ 01].$$

Weighing the basis vectors with respect to the obtained coefficients, we obtain expressions of the logic functions in the given basis

$$(x_1 \& x_2) \oplus 1 = 1 - (x_1 + x_1 - 2x_1 \& x_1);$$

$$min(x_1, x_2) \oplus 1 = 1 + 3min\left(x_1^{(1)}, x_2^{(1)}\right) - 2min\left(x_1^{(2)}, x_2^{(2)}\right);$$

$$min(x_1, x_2) \ominus 1 = 2 - 3min\left(x_1^{(1)}, x_2^{(1)}\right) + min\left(x_1^{(2)}, x_2^{(2)}\right).$$

The functional schemes corresponding to these expressions are shown in Figure 7 *a, b, c.*

To obtain the equivalent expressions of the functions in linear analogs of the logical-arithmetic basis under consideration, one can proceed in two ways:

- replace the used basis vectors with their linear analogs in the expressions obtained;

- use the above relations between logic and linear operations and convert the expressions written above into the linear form.

In all three cases, we obtain the following results:

$$(x_1 \& x_2) \oplus 1 = 1 - \{x_1 + x_1 - 2[(x_1 + x_1) \div 1]\};$$

$$min(x_1, x_2) \oplus 1 = 1 + 3\{[1 \div (1 \div x_1)] \div (1 \div x_2)\} - -2\{[(x_1 \div 1) + (x_2 \div 1)] \div 1\};$$

$$min(x_1, x_2) \ominus 1 = 2 - 3\{[1 \div (1 \div x_1)] \div (1 \div x_2)\} + +\{[(x_1 \div 1) + (x_2 \div 1)] \div 1\}.$$

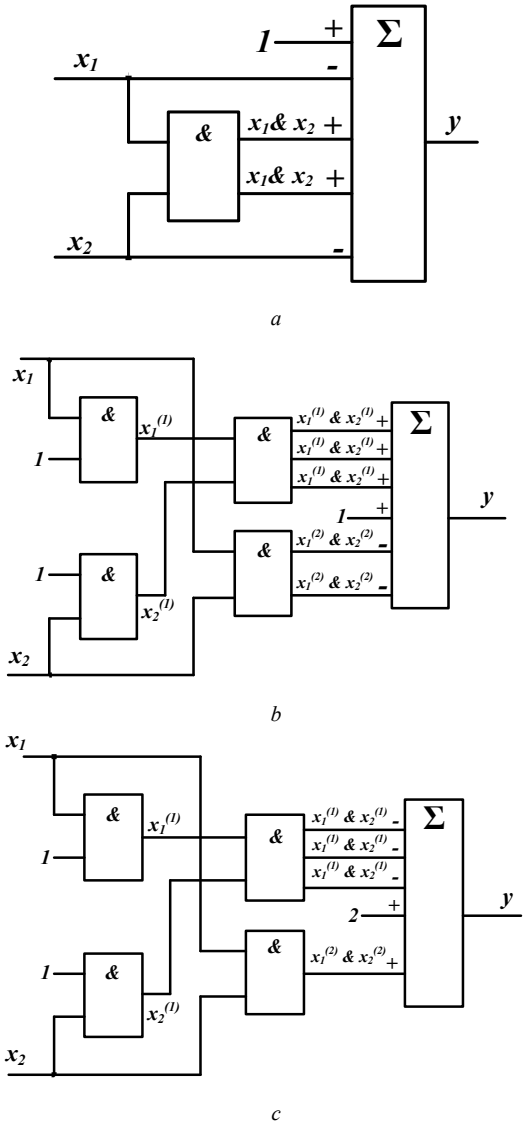


Figure 7: Functional schemes of the logical-arithmetic realization of the circular shift operation: a - two-valued, b - three-valued with a shift to the right, c - three-valued with a shift to the left.

The functional schemes corresponding to these expressions are shown in Figure 8, *a, b, c.*

The difference between the linear representations obtained from the arithmetic-logical representation is the possibility of physical realization. The authors of the article have obtained more than 25 patents for the circuit implementation of two-valued and three-valued logic circuits.

D. Expansion of the arbitrary function into the algebraic sum of monotonic functions of the same value. To obtain the representation of the functions under consideration by the algebraic sum of monotonic functions, we expand them into the algebraic sum of monotonic functions within each value. For the two-valued function, the sequence graph of nondecreasing components has the form shown in Figure 9.

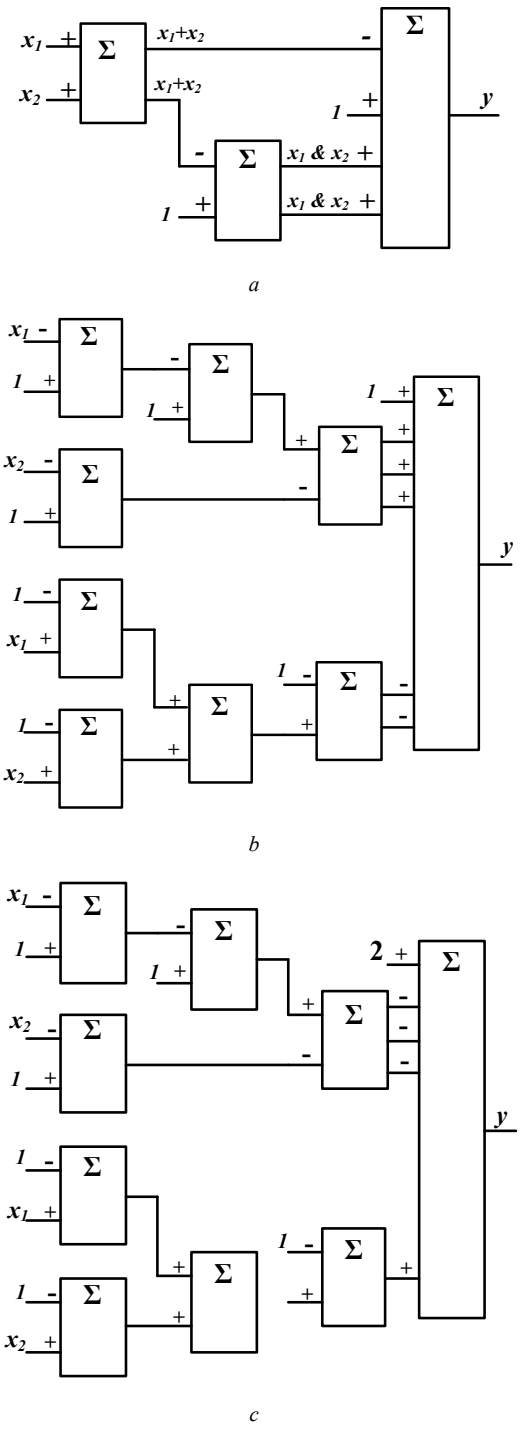


Figure 8: Functional schemes of the linear realization of the circular shift operation based on the *truncated difference*: *a* - two-valued, *b* - three-valued with a shift to the right, *c* - three-valued with a shift to the left.

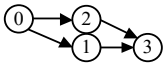


Figure 9: Sequence graph of nondecreasing components of the two-valued functions of two variables.

The structure of the expansion function of the equivalence $\bar{x}_1\bar{x}_2 \vee x_1x_2 = (x_1\&x_2) \oplus 1$ into the algebraic sum of monotonic functions is given in Figure 10.

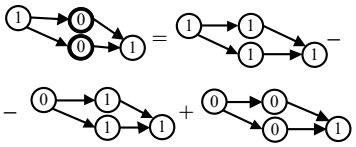


Figure 10: The structure of the expansion of the two-valued function $\bar{x}_1\bar{x}_2 \vee x_1x_2$ into the algebraic sum of monotonic functions.

Here and below, the double circle marks the terms the values of which violate the monotonicity of the function.

It follows from the structure of the expansion that the representation of the function under consideration by the algebraic sum of monotonic functions has the following form

$$f(\tilde{x}^{(2)}) = (x_1\&x_2) \oplus 1 = 1 - x_1 \vee x_2 + x_1\&x_2.$$

Carrying out the similar transformations for the three-valued functions, the sequence graph of nondecreasing components of which is shown in Figure 11.

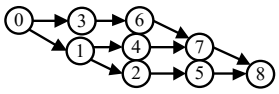


Figure 11: Sequence graph of nondecreasing components of the three-valued functions of two variables.

we obtain the following representations of the functions under consideration by the algebraic sum of monotonic functions:

- for $min(x_1, x_2)\oplus 1$.

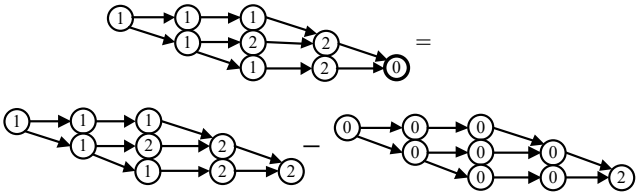


Figure 12: The structure of the expansion of the three-valued function $min(x_1, x_2)\oplus 1$ into the algebraic sum of monotonic functions.

$$min(x_1, x_2)\oplus 1 = 1 - min\left(x_1^{(1)}, x_2^{(1)}\right) -$$
$$-2min\left(x_1^{(2)}, x_2^{(2)}\right) + min\left(x_1^{(1)}, x_2^{(1)}\right);$$

– for $min(x_1, x_2) \ominus 1$.

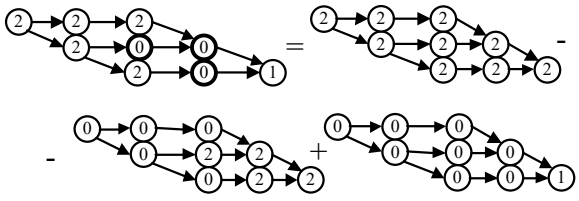


Figure 13: The structure of the expansion of the three-valued function $min(x_1, x_2) \ominus 1$ into the algebraic sum of monotonic functions.

$$min(x_1, x_2) \ominus 1 = 2 - 2min\left(x_1^{(1)}, x_2^{(1)}\right) +$$
$$+min\left(x_1^{(2)}, x_2^{(2)}\right).$$

To obtain the equivalent representations of the expansion functions in linear analogs of the basis under consideration, we can proceed in the same way as with the functions themselves.

Expansion of the arbitrary function into the algebraic sum of smaller value.

We consider this operation with the help of the above transformation of the three-valued function $min(x_1, x_2)\oplus 1$, by unitary and weighted coding of the values of variables and functions.

In these cases, the encoding of variables and functions looks like this:

- unitary coding

x	x_2	x_1
0	0	0
1	0	1
2	1	1

Hence, it follows that for the unitary coding

$x = x_1 + x_2;$

– weighted coding

x	x_2	x_1
0	0	0
1	0	1
2	1	0

It follows that at the weighted coding

$$x = x_1 + 2x_2.$$

Using the basic and its inverse matrix, we obtain the expressions for the two-valued expansion functions from the multivalued arguments:

- for the unitary coding

$$f_1 = 1 + min\left(x_1^{(1)}, x_2^{(1)}\right) - min\left(x_1^{(2)}, x_2^{(2)}\right);$$

$$f_2 = 2min\left(x_1^{(1)}, x_2^{(1)}\right) - min\left(x_1^{(2)}, x_2^{(2)}\right);$$

$$f = f_1 + f_2;$$

- for the weighted coding

$$f_1 = 1 - min\left(x_1^{(1)}, x_2^{(1)}\right);$$

$$f_2 = 2min\left(x_1^{(1)}, x_2^{(1)}\right) - min\left(x_1^{(2)}, x_2^{(2)}\right);$$

$$f = f_1 + 2f_2.$$

The representation of the three-valued function $min(x_1, x_2)\oplus 1$ by the two-valued ones for both versions of coding is given in Table 2.

Table 2 – The expansion of the three-valued function $min(x_1, x_2)\oplus 1$ into the sum of the two-valued ones

x_2	x_1	Unitary		Weighted	
		f_2	f_1	f_2	f_1
0	0	0	1	0	1
0	1	0	1	0	1
0	2	0	1	0	1
1	0	0	1	0	1
1	1	1	1	1	0
1	2	1	1	1	0
2	0	0	1	0	1
2	1	1	1	1	0
2	2	0	0	0	0

Replacing the three-valued variables with the two-valued ones in accordance with the rules given above, we obtain the two-valued expansion functions in the following form:

- for the unitary coding

$$f_1 = 1 - x_{11}x_{12}x_{21}x_{22};$$

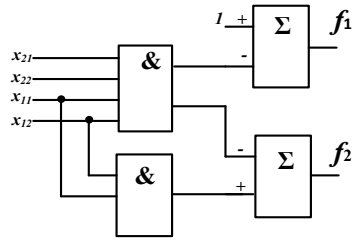
$$f_2 = x_{11}x_{21} - x_{11}x_{12}x_{21}x_{22};$$

- for the weighted coding

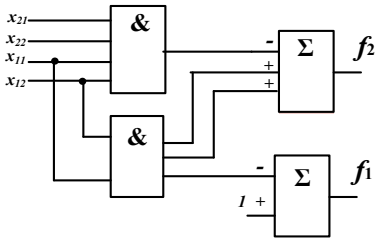
$$f_1 = 1 - x_{11}x_{21};$$

$$f_2 = 2x_{11}x_{21} - x_{11}x_{12}x_{21}x_{22}.$$

The functional schemes of the linear realization of the circular shift operation by the expansion into the sum of the two-valued functions are shown in Figure 14.



a



b

Figure 14: Functional schemes of linear realization of the circular shift operation by the expansion into the sum of the two-valued functions: *a* - on the basis of the unitary coding, *b* - on the basis of the weighted coding.

For the threshold representation and the subsequent threshold realization, it suffices to transform the equations obtained above into the threshold form. After completing this transformation, we get:

- for the unitary coding

$$f_1 = 1 > [(x_{11}+x_{12}+x_{21}+x_{22}) > 3];$$

$$f_2 = [(x_{11}+x_{21}) > 1] - [(x_{11}+x_{12}+x_{21}+x_{22}) > 3];$$

- for the weighted coding

$$f_1 = 1 - [(x_{11}+x_{21}) > 1];$$

$$f_2 = 2[(x_{11}+x_{21}) > 1] - [(x_{11}+x_{12}+x_{21}+x_{22}) > 3].$$

The functional schemes shown in Figure 15 correspond to these equations.

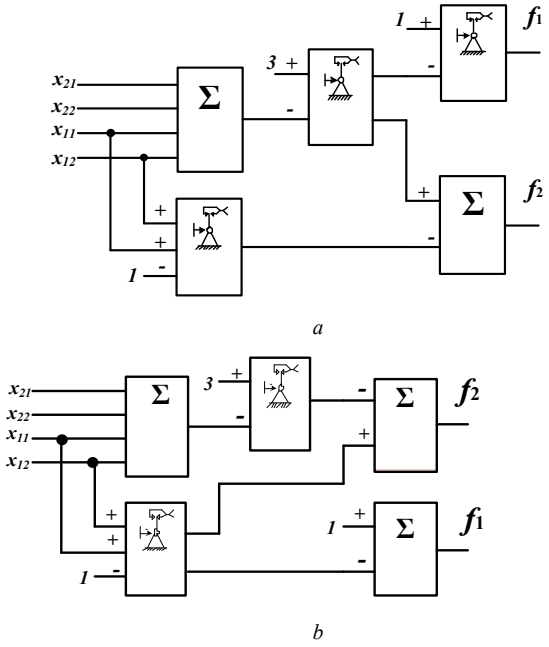


Figure 15: Functional schemes of linear realization of the circular shift operation by the expansion into the sum of the threshold two-valued functions: *a* - on the basis of the unitary coding, *b* - on the basis of the weighted coding.

The considered complex of transformations of logic functions in linear algebra proves useful in the design of digital structures for various applications.

5. Circuitry of the Linear Logic and Threshold Elements

A. Basic functional nodes of digital logic elements. The implementation of the mathematically predetermined set of linear operations over the current signals requires the corresponding set of hardware. The circuit implementation of digital signal transformation functions based on the mathematical tool of linear algebra can be reduced to performing a relatively simple set of operations over the current signals. These operations include:

- conversion of the standard logic signals into the binary current signals matched with a reference current quantum I_0;

- formation of the multivalued (non-binary) algebraic sums of current signals;

- performing the *comparing* operations of the received sums with the given levels of the reference currents.

These operations are typical for analog microcircuitry, therefore the main nodes of various functional elements can be constructed on the basis of the widely used analog structures. In addition, TTL circuitry and CMOS circuitry of these operations completely coincide.

The reference signal driver. It is designed to generate voltages that provide operational modes of elements of digital circuits synthesized in linear algebra. The schematic configuration of the driver is shown in Figure 16.

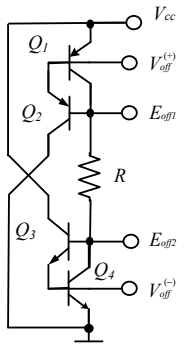


Figure 16: Reference signal driver.

The symmetric structure of the reference signal source is necessary for realizing the operations of converting "inflowing" and "flowing out" currents when creating the algebraic sums of variables in the mathematical tool of linear algebra. The only current-stabilizing two-terminal network, in particular case the resistor *R*, determines the levels of all reference signals. $V_{off}^{(+)}$ and $V_{off}^{(-)}$ are reference for setting the operating modes of current mirrors, and E_{off1} and E_{off2} - for setting the offset in the comparators of the currents. It can also be replaced by some semiconductor structure, and then the circuit becomes completely homogeneous and highly technological. It is also possible to make the two-terminal network *R* as an external element, which will allow changing the power consumption and the associated characteristics of the LSI during the debugging process.

The reference current source (RCS, Figure 17). The operating mode of the RCS is set by the reference voltages $V_{off}^{(+)}$ for the current sources and $V_{off}^{(-)}$ for the current sinks. The problem of constructing the RCS for digital circuits synthesized in linear algebra is completely similar to their problems in analog circuitry.

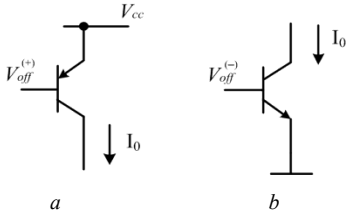


Figure 17: RCS of: a – the current current, b – the current sink.

The current follower (logic interpretation - direction converter). It is intended for the coordination of current directions at their algebraic summation. The schematic configurations of some versions of current followers in TTL circuitry are shown in Figure 18.

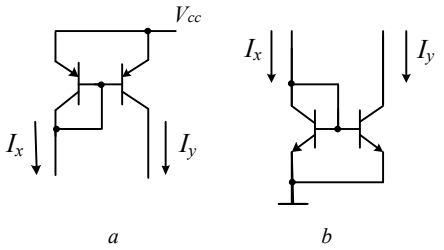


Figure 18: Current direction converter of: *a* – the sink current, *b* - the source current.

Circuitry of CMOS converters of the current direction is similar.

Algebraic current adder. It is an wired pack of the outputs of current mirrors with current directions determined by the mathematical representation of the realized logic function. To agree the operating modes of subsequent elements, it is provided with a buffer stage. The schematic configuration of the algebraic adder is shown in Figure 19.

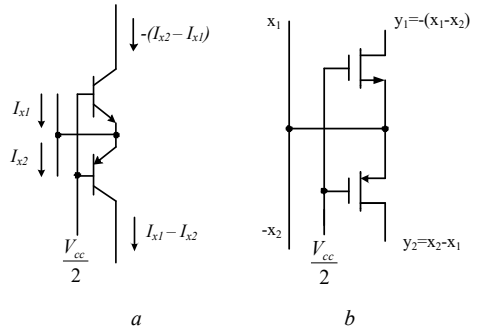


Figure 19: Algebraic current adder:
a - TTL circuit, *b* - CMOS circuit.

Current comparators. They are designed to determine the excess value of one of the input currents over the other. The schematic configurations of current comparators are given in Figures 20 and 21.

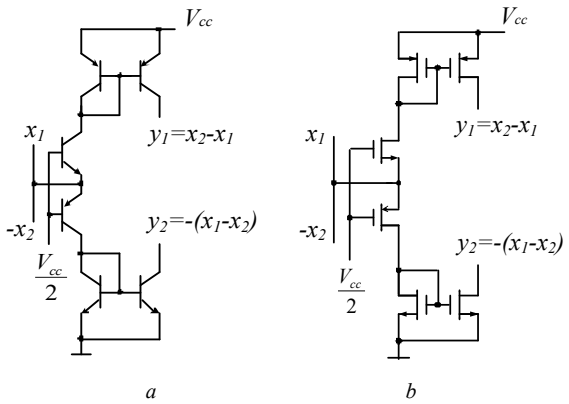


Figure 20: Comparator based on the *truncated difference*:
a - bipolar circuit, *b* - CMOS circuit.

In the circuit in Figure 21, the excess of one of the input currents (I_{x1}) over the other (I_{x2}) is determined by subtracting the second current from the first. One of the compared currents must be a source current and the second one – a current sink. Such a comparator is applicable for any logic value.

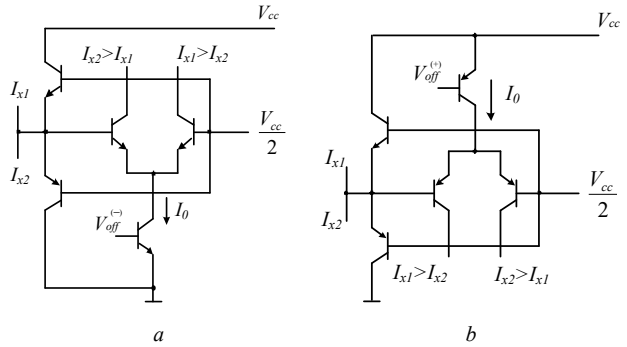


Figure 21: Current comparator based on the comparison: *a* - for current sinks, *b* - for source currents.

In the circuit of Figure 21, *a* at $I_{x1} > I_{x2}$ the left-hand transistor of the differential stage (DS) is closed and the current of the current mirror "leaves" through the right transistor of the DS. At $I_{x1} < I_{x2}$, the right transistor of the DS is closed and the current of the current mirror "goes" through the left transistor. The output current of the DS I_0 is the current sink.

In the circuit in Figure 21, *b* everything appears in a similar way, but with other current directions. Such a scheme is applicable for implementation of the two-valued digital structures.

In the previous section, the difference between the two-valued and multivalued implementations of the *comparison* operation is shown. In the latter case, it is possible to determine not only the fact, but the magnitude of the excess of one input signal over another one. In Figure 22 there is a circuit of the three-valued comparator performing such a modified *comparison* operation.

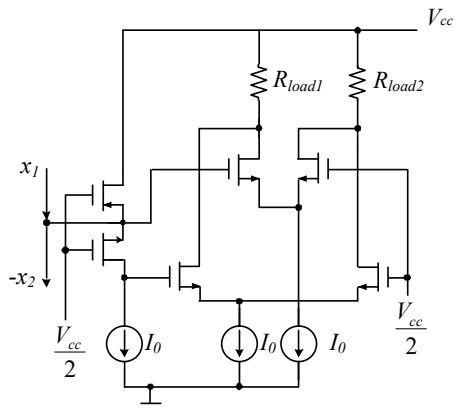


Figure 22: Modified multivalued comparator based on the comparison.

Such a scheme can be constructed for any value, increasing accordingly the number of parallelly operating DSs available to work with different values of currents.

Logic elements AND, OR, NOT. The representations of the operations of the two-valued basic functionally complete system in linear algebra using the *truncated difference* have the following form

$$x_1 \& x_2 = (x_1 + x_2) \div 1;$$

$$x_1 \vee x_2 = 1 \div [1 \div (x_1 + x_2)];$$

$$\bar{x} = 1 - x.$$

The schematic configuration of these elements are shown in Figures 23, 24 and 25, correspondingly.

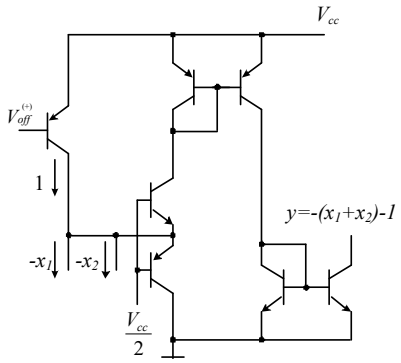


Figure 23: Schematic configuration of the element AND.

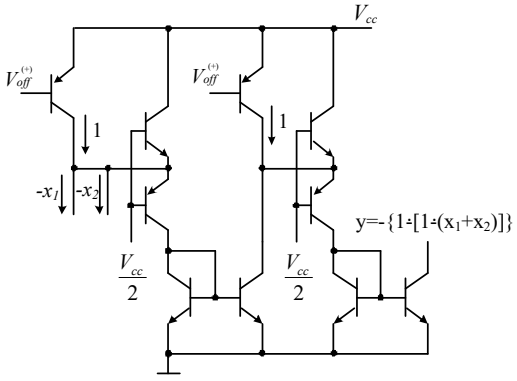


Figure 24: Schematic configuration of the element OR.

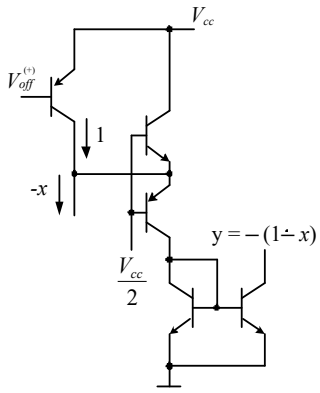


Figure 25: Schematic configuration of the element NOT.

Comparing the schemes shown in Figures 24 and 25, it is easy to see that removing the circuit of in Figure 25 from the scheme of Figure 24, we obtain the OR-NOT element.

Logic element "Inhibition". The representation of the two-valued operation in linear algebra using the *truncated difference* has the following form

$$x_1 \& \bar{x}_2 = x_1 \div x_2.$$

For the multivalued version, the two-valued inversion operation (in accordance with the accepted generalization ideology) should be replaced with the direct $min(x_1, x_2 \oplus 1)$, or inverse min $min(x_1, x_2 \ominus 1)$ cycle operation, or left unchanged; i.e. in the following form $min(x_1, 1 \div x_2)$. Other ideologies of generalization are also possible.

When using the *comparison* operation, the above expressions for operations of the functionally complete system take the following form

$$x_1 \& x_2 = (x_1 + x_2) > 1;$$

$$x_1 \vee x_2 = (x_1 + x_2) > 0;$$

$$\bar{x} = 1 > x.$$

The implementation of these functions on the basis of the *comparison* operation can be performed with the help of the universal logic element (ULE), the schematic configuration of which is shown in Figure 26.

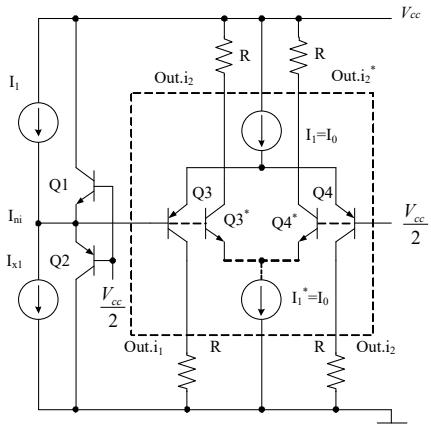


Figure 26: Basic scheme of the universal logic element.

At the inequality $I_1 < I_{x1}$, the source difference current is generated at the node In.i of the ULE. It will "go" to the emitter of the transistor Q1, increasing the voltage at the first input (In.1) of the voltage comparator up to the value $V_{OFF} + U_{be1}$, where $U_{be1} \approx 0.7V$ - voltage of the open emitter junction of the transistor Q1. In this case, the input transistors Q3 and Q4 of the voltage comparator (VC) switch to the inverse states - the collector current of the transistor Q3 becomes zero, and the transistor Q4 starts to transmit the current of the reference current source $I_2 = I_0$ to the second current output (Out.i_2) of the ULE.

Thus, depending on the difference in the numerical values of the currents I_1 and I_{x1}, the output currents of the ULE take one of two values: either it is the current of the reference current source $I_1 = I_0$ or "zero" (no current). Since the current I_1 (I_1^*) is equal to the current quantum I_0, then in one of the current outputs of the voltage comparator a standard current signal I_0 of one of the logic levels is generated, and in its second output - an inverse logic level signal. Depending on the numerical values of I_1 ($I_1 = 0{,}5I_0$, $I_1 = I_0$, $I_1 = 1{,}5I_0$) and the methods of forming the input current signals of the ULE (Figure 26), various logic functions can be performed, for example

– $\bar{x} = 1 - x$ at $y_1 = \begin{cases} I_0 \; at \; I_1 > I_x \\ 0 \; at \; I_1 \le I_x \end{cases}$;

– $x_1 \& x_2 = (x_1 + x_2) > 1$ when $y_1 = \begin{cases} I_0 \; at \; 1{,}5I_1 > I_x \\ 0 \; at \; 1{,}5I_1 \le I_x \end{cases}$;

– $x_1 \vee x_2 = (x_1 + x_2) > 0$ if $y_1 = \begin{cases} I_0 \; at \; 0{,}5I_1 > I_x \\ 0 \; at \; 0{,}5I_1 \le I_x \end{cases}$.

In the last two expressions $I_x \equiv x_1 + x_2$.

The element considered above can be used as a logic element, or as a threshold one; i.e. the generalization of the ordinary and threshold logics in the linear representation is very close. In the first case, unitary variables are fed to the inputs of the element. Besides, the number of inputs must correspond to the number of the variables. In the second case, the weighted sum of the variables should be fed to the "positive" input, and the constant equal to the calculated threshold value should be fed to the "negative" input. It is noteworthy that the described element can serve as an element of the homogeneous matrix, which can be used for matrix synthesis of the current digital structures. In addition, it is possible to construct universal current logic modules on its basis.

Similarly, the Ban operation can be expressed in terms of the *truncated difference* as

$$min(x_1, \{x_2 + 1 \dot{-} k[1 \dot{-} (x_2 \dot{-} 1)]\});$$

$$min(x_1, \{x_2 - 1 + k[1 \dot{-} x_2]\}),$$

and through the *comparison* - as

$$min(x_1, [x_2 + 1 \dot{-} k(x_2 > 2)]);$$

$$min(x_1, \{x_2 - 1 + k[1 > x_2]\}),$$

Cut formers. In fact, the cut former is an input signal limiter at the given level. As a former, the schemes that implement the operations $min(x_1, x_2)$ (upper cuts) and $max(x_1, x_2)$ (lower cuts) can be used.

Buffer output stage. The combination of the algebraic adder and the current follower (converter) can be applied as a buffer output stage. With the help of the latter, the given number of the output circuits can be arranged to provide the required output branching factor.

B. The synthesis of logical schemes. Let's consider it through the example of the two-valued and multivalued circular shift elements discussed above. They can be represented as a *single operation* in the form $min(x_1, x_2 \oplus 1)$ or $m(x_1, x_2 \ominus 1)$, or as a *compound operation*, i.e. as the sequential combination of operations working on one another $min(x_1, x_2)$ or $max(x_1, x_2)$ and operations $\oplus 1$ or $\ominus 1$, that is, in the form $min(x_1, x_2) \oplus 1$ or $min(x_1, x_2) \ominus 1$. We confine ourselves to the synthesis of the circuits using the operation $min(x_1, x_2)$.

In accordance with [17-32], the synthesis of the functional scheme corresponding to any logic function consists in multiplying the row-vector of the values of the function by the inverse basic matrix and obtaining the expansion vector of the function from the selected basis, and then recording the representation of the function as a weighted sum of the basis vectors. We perform the synthesis of the selected schemes using the basis presented below.

In the two-valued case, the element AND-NOT has an arithmetic-logical representation, described by the expression

$$\overline{x_1 \& x_2} = 1 - x_1 \& x_2,$$

which can be represented by the *truncated difference* in the following form

$$\overline{x_1 \& x_2} = 1 - [(x_1 + x_2) \dot{-} 1],$$

and by *comparison* - in the following form

$$\overline{x_1 \& x_2} = 1 > [(x_1 + x_2) > 1].$$

The schematic configurations of the elements can be constructed directly according to these expressions. Figure 27 shows an element scheme based on the *truncated difference*.

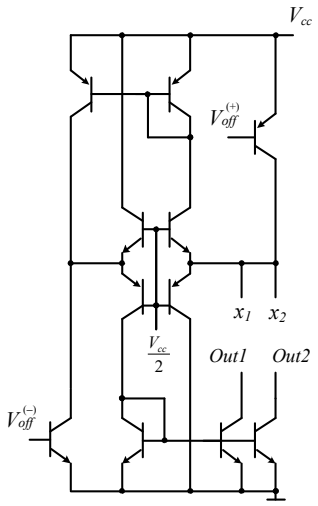


Figure 27: Schematic configuration of the element of AND-NOT based on the *truncated difference.*

We proceed similarly for the three-valued schemes in the first case (by the single operation). Using the basis

$$min(B1, B1)^{(3)} = \begin{bmatrix} 1 \\ 1 \dot{-} (1 \dot{-} x_1) \\ x_1 \\ 1 \dot{-} (1 \dot{-} x_2) \\ [1 - (1 \dot{-} x_1)] - (1 \dot{-} x_2) \\ x_1 \dot{-} 2(1 \dot{-} x_2) \\ x_2 \\ 2[1 - (1 \dot{-} x_1)] - (2 \dot{-} x_2) \\ x_1 \dot{-} (x_1 \dot{-} x_2) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 2 \end{bmatrix}$$

with the inverse matrix

$$\left(min(B1,B1)^{(3)}\right)^{-1} =$$

$$= \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 & -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix},$$

we obtain the following results for the *single operation*:

- resolution vectors

$$w(min(x_1,x_2)\oplus 1) = [1\ 0\ 0\ 0\ 3\ 0\ 0\ 0\ -2];$$

$$w(min(x_1,x_2) \ominus 1) = [2\ 0\ 0\ 0\ -3\ 0\ 0\ 0\ 1];$$

- linear expressions of the functions

$$min(x_1,x_2) \oplus 1 = 1 + 3[1 - (1 \div x_1) - (1 \div x_2)] -$$

$$-2[x_1 \div (x_1 \div x_2)] .$$

$$min(x_1,x_2) \ominus 1 = 2 - 3[1 - (1 \div x_1) - (1 \div x_2)] +$$

$$+[x_1 \div (x_1 \div x_2)].$$

The schematic configuration of the element of the right circular shift, synthesized directly from the above expression, is shown in Figure 28.

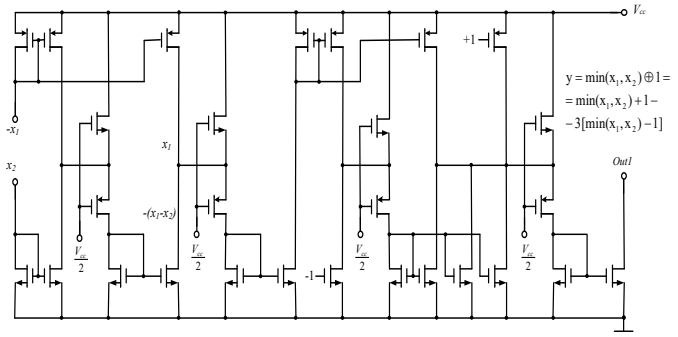


Figure 28: CMOS element of the right circular shift.

The schematic configuration of the element of the left circular shift, synthesized directly from the expression given above, is given in Figure 29.

Similarly, for the *compound operation*:

- the resolution vector $min(x_1,x_2)$

$$w\big(min(x_1,x_2)\big) = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1],$$

- linear expression of the function $min(x_1,x_2)$

$$min(x_1,x_2) = x_1 \div (x_1 \div x_2). \quad (5)$$

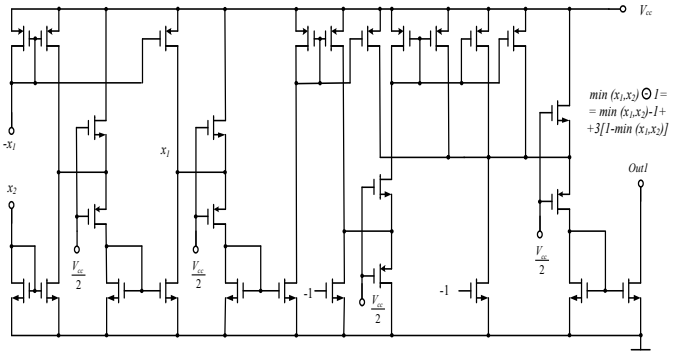


Figure 29: CMOS element of the left circular shift.

Similarly, for the *compound operation*:

- the resolution vector $min(x_1,x_2)$

$$w\big(min(x_1,x_2)\big) = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1],$$

- linear expression of the function $min(x_1,x_2)$

$$min(x_1,x_2) = x_1 \div (x_1 \div x_2). \quad (5)$$

- resolution vectors of the operations $x\oplus 1$ or $x \ominus 1$:

$$w(x\oplus 1) = [1\ 3\ -2];$$

$$w(x \ominus 1) = [2\ -3\ 1].$$

- linear expressions of the functions:

$$x \oplus 1 = 1 + [1 \div (1 \div x)] - 2x;$$

$$x \ominus 1 = 2 - 3[1 \div (1 \div x)] + x.$$

The schematic configuration of the element $min(x_1,x_2)$, synthesized directly from expression (5) discussed above, is shown in Figure 30.

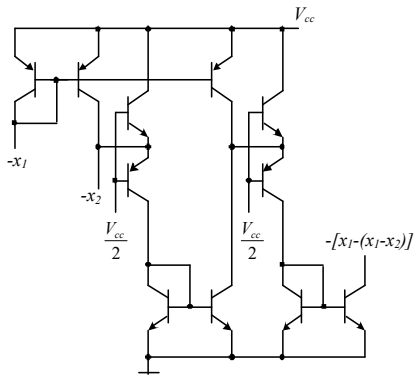


Figure 30: Schematic configuration of the element $min(x_1,x_2)$.

The schemes of the elements of the left and right cycles are shown in Figures 31 and 32.

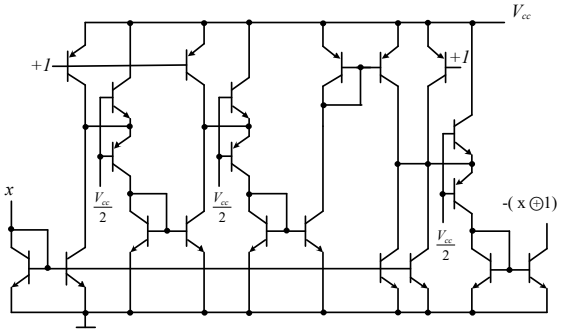


Figure 31: Element of the right cycle.

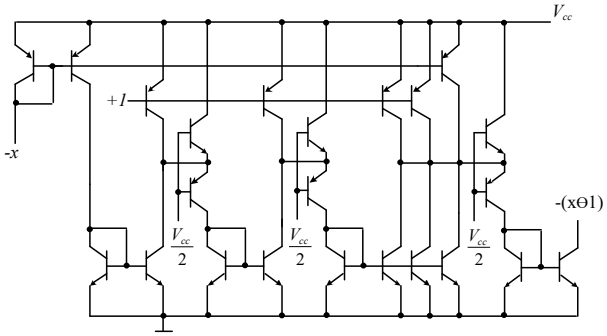


Figure 32: The element of the left cycle.

To obtain the final expression in the procedure for synthesizing a particular logic element, it is sufficient to substitute $min(x_1, x_2)$, for x in the last expression, and for the circuit implementation, - use the above implementation of the function $min(x_1, x_2)$, then connect the circuit $(x \oplus 1)$ or $(x \ominus 1)$ to its output.

C. To demonstrate the possibilities of the *sequential circuit synthesis*, we use the results of the logic element synthesis obtained above.

Boolean approach to the logic synthesis of triggers consists in supplying the memory element with a control circuit that provides a specified law for the operation of a specific type of the trigger. In the two-valued case, the memory element is a scheme of two 2AND-NOT elements covered by the positive feedback (Figure 33):

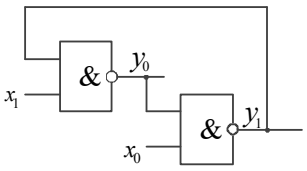


Figure 33: Two-valued trigger (memory element).

The linear synthesis of the two-valued triggers in linear algebra does not differ fundamentally from Boolean synthesis [17]. The schematic configuration of the trigger can be constructed in the same way as it is done in Boolean logic (for example – Figure 34):

When moving to higher values while remaining the general idea of synthesis, it is necessary to use the elements and operations that are a generalization of the two-valued operations and logic elements. The operations *min* (x_1,x_2) and *max* (x_1,x_2) are generalization of the operations & and ∨ is. As for the inversion operation, to generalize it for a multi-valued case it is convenient to represent it in the form of $\bar{x} = x \oplus 1 = x \ominus 1$ and generalize it by a circular shift of Post (left or right). The general functional configuration of the memory element based on three-valued elements of the direct circular shift is shown in Figure 35.

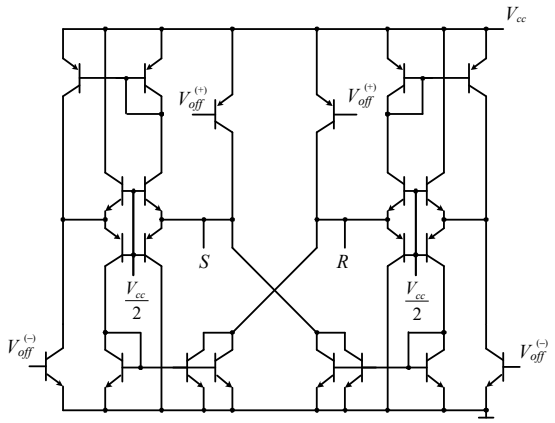


Figure 34: Two-valued RS-flip-flop in the studied basis based on the *truncated difference.*

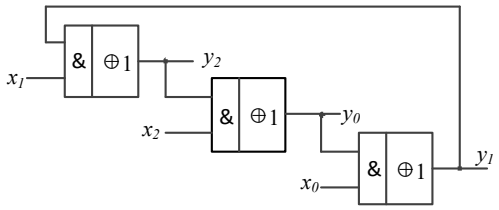


Figure 35: Three-valued trigger on the elements of the direct circular shift.

The general functional configuration of the memory element based on the three-valued elements of the inverse circular shift is given in Figure 36.

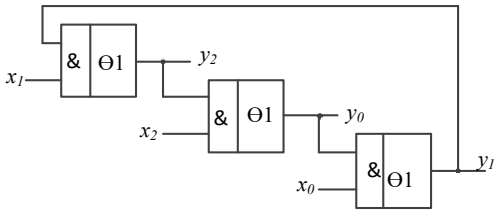


Figure 36: Three-valued trigger on the elements of the inverse circular shift.

The functional schemes of the memory elements of higher significance look similar.

Thus, to construct any multivalued memory element, it is necessary to synthesize the elements of direct and inverse circular shifts. To create a trigger of the given type (D-, RS-, JK-, etc.), it is required to equip the memory element with the corresponding control circuit.

The control circuits of the memory elements are constructed on the base of the verbal description of the operation of a specific type of the trigger. The result of the analysis of this description is the detection of logic control functions for each input of the memory element.

For example, the RS-flip-flop operates according to the following algorithm:

- the values of the input signals $S = R = k-1$ correspond to the storage mode;

- the signal *S* increases the trigger state index relatively to the current state towards the state *k*-1, and the signal R reduces the state index towards the state "0";

- the value of the state index change is equal to the value of the input signal: the signal equal to 1 can increase (or decrease) the current state index by 1, the signal equal to 2 - by 2, etc .; besides, the signal S increases the trigger state index to the state *k*-1, and the signal *R* reduces the state index to "0"; the state change on the cycle is impossible (we leave this for the universal triggers!);

- to change the trigger state index relatively to the current state *i*, the signal *S* can take values from 1 to *k*-1-*i* (larger values are equivalent to the value of *k*-1-*i*). Similarly, the signal *R* can take the values from 1 to *i* -1 (larger values are equivalent to the value *i* -1);

- all combinations are inhibited combinations of values of the input signals, except 0 – *k*-1, 1 – *k*-1, *k*-2 – *k*-1, *k*-1 – *k*-1, *k*-1 – 0, *k*-1 – 1, …, *k*-1– *k*-2.

Now it is possible to synthesize the logic function of the control circuit of the *RS*-flip-flop. Its output signal is a multi-valued signal *x* that sets the next state of the memory element, and the input signals are multivalued signals *S* and *R*, as well as the current trigger state index.

The truth table of the control functions of the three-valued RS-flip-flop, compiled on the basis of this description, has the following form:

S	R	Q_{t+1}	x_0	x_1	x_2
0	2	Q_2	0	1	1
1	2	Q_1	1	0	1
2	2	Q_t	1	1	1
2	1	Q_1	1	0	1
2	0	Q_0	1	1	0

From the last table it follows that the logic functions for controlling the state of the memory element are described as follows:

$$x_0 = P(S > 0) = 1 \div (1 \div S);$$

$$x_1 = 1 > [1 > (1 > S)] > [1 > (1 > R)] =$$

$$= 1 \div [1 \div |1 \div S|] \div [1 \div |1 \div R|];$$

$$x_2 = P(R > 0) = 1 \div (1 \div R).$$

The combination of the memory element circuit and the control circuit results in the implementation of the trigger of the given type.

Similarly, arguing, we can obtain the logical control functions of the three-valued D-flip-flop:

$$x_0 = (1 \div C)\&(1 \div D) = 1 \div [(1 \div D) \div C] =$$

$$= 1 \div [(1 \div C) \div D];$$

$$x_1 = (1 \div C)\&[D \div 2(D \div 1)] = 1 \div \{D \div [2(D \div 1) \div C]\} =$$

$$1 \div [1 \div (|D \div 1| + C)];$$

$$x_2 = (1 \div C)\&(D \div 1) = 1 \div [(D \div 1) \div C] =$$

$$= D \div [(1 \div C) \div 1].$$

In the same way, we can obtain the description of the control system of any other multivalued trigger, the number of types of which is certainly greater than the two-valued one.

Combining the control circuit and the memory element in series, it is possible to obtain the schematic configuration of the trigger of any type and any value.

6. Conclusion

1. The mathematical tool of linear algebra can be newly applied in problems of logic synthesis and circuit implementation of the current digital structures.

2. Linear algebra:

- enables to create not only a two-valued, but also a really properly functioning multivalued element base for digital signal processing devices;

- conduces the design of the digital element base with the improved technological, technical and operational characteristics (in comparison with the potential logic based on Boolean algebra);

- can serve as a basis for creating LSI on the basis of matrix fields of homogeneous elements (as in modern Altera design systems, etc.);

- improves the reliability of current digital LSIs synthesized on its basis, under extreme operating conditions (temperature, radiation, in-phase interferences, etc.).

3. Two-valued and three-valued triggers based on cyclic shift elements are considered. It is shown that in the transition to greater significance, while preserving the general idea of synthesis, it is necessary to use operations that are generalizations of two-valued operations and the corresponding logical elements.

4. The presents a basic set of current logic elements for the devices of automation, which allows solving the problems of transformation of the current signals in a different and more efficient way.

5. In the schemes of the developed class is provided a differential representation of the output signal that minimizes the effect of temperature and radiation on their basic parameters.

Conflict of Interest

The authors declare that there is no conflict of interests regarding of publication of this paper.

Acknowledgment

The research is carried out at the expense of the Grant of the Russian Science Foundation (project № 16-19-00122).

References

[1] N. N. Prokopenko, N. I. Chernov, V. Ya. Yugai, N. V. Butyrlagin, "The Element Base of the Multivalued Threshold Logic for the Automation and Control Digital Devices," on International Siberian Conference on Control and Communications, SIBCON-2017, Astana, Kazakhstan, 29-30 June, 2017.

[2] A. S. Karpenko, "Multi-valued logics (monograph)," in series "Logics and computer", Moscow, issue 4, 1997, 223 p. (in Russian).

[3] L. A. Zalmanson, "Fourier transform, Walsh, Haar and their application in management, communications and other fields," Moscow, Nauka, Fizmatlit publ., 1989, 496 p. (in Russian).

[4] V. S. Vykhovanets, V. D. Malugin, "Spectral methods in logical management," Proceedings of the 2nd international scientific-technical conference "Modern methods of digital processing of signals in spelthorne, monitoring, diagnosis and control (OS 1998)", Minsk, 1998, pp. 56-59. (in Russian).

[5] M. Thornton, "Mixed representation of the multi-Boolean function spectra and critical charts," Automatics and telemechanics, 2004, vol.6, pp. 188-200. (in Russian).

[6] G. I. Ivchenko, V. A. Mironov, "Some questions of spectral analysis of random Boolean functions with constraints," Discrete mathematics, 2013, vol. 1, pp. 90-110. (in Russian).

[7] V.N. Kondratyev, A.A. Shalyto, "Implementation of systems of Boolean functions using arithmetical polynomials," Automatics and telemechanics, 1993, vol. 2, pp. 114-122. (in Russian).

[8] V. D. Malyugin, "The parallel logic computation by the arithmetic of polynomials," Moscow, Nauka, Fizmatlit publ., 1997, 192 p. (in Russian).

[9] A. V. Sokolov, O. N. Zhdanov, A. A. Ayvazyan, "Methods for the synthesis of the algebraic normal forms of functions of multivalued logic," System analysis and applied Informatics, No. 1, 2016, p. 69-76. (in Russian).

[10] M. Dertouzos, "Threshold logic," Moscow: Mir publ., 1967. (in Russian).

[11] V. G. Nikonov, "Threshold representations of Boolean functions," Review in applied and industrial mathematics, Series discrete mathematics, 1994, vol. 1, No. 3, pp. 402-457. (in Russian).

[12] E. A. Butakov, "Methods of synthesis of relay devices," Moscow: Energy publ., 1970, 328 p. (in Russian).

[13] E. N. Vavilov, etc. The Synthesis of threshold circuits for the elements – Moscow: Soviet Radio publ., 1970, 368 p. (in Russian).

[14] S. Muroga, "Threshold logic and its applications," New York: Wiley, 1971.

[15] W. H. Piers, "Failure Tolerant computer dezign," New York and London: Academic Press, 1965.

[16] V. G. Nikonov, N. Nikonov, "Features of threshold concepts k-valued functions," Tr. on discr. mod., 2008, volume 11, issue 1, pp. 60-85. (in Russian).

[17] J. Hastad, "On the size of weights for threshold gates." SIAM J. Discr. Math. 1994.

[18] N. I. Chernov, "Foundations of Logic Synthesis of Real Numbers Field Digital Structures", Taganrog: TRTU, 2000, p.146 (in Russian).

[19] N. I. Chernov, "Boolean Linear Space as an Algebraic Structure for Logic Synthesis of Digital Units," The News of TRTU, 2003, No.1, pp. 215-220. (in Russian).

[20] N. I. Chernov, "Structural Synthesis of Digital Units within Boolean Linear Spaces," The News of TRTU, 2003, No.2, pp. 73-76. (in Russian).

[21] N. I. Chernov, "Efficiency of Application of the Body of Linear Spaces within Logic Synthesis," Conferences on Artificial Intelligence Systems (IEEE AIS' 05)' and Intelligent CAD (CAD-2005, vol.1, pp. 420–424. (in Russian).

[22] N. I. Chernov, "Logic Design of Digital Structures on Controlled Current Generator," The News of TREU, 2005, No.11, pp. 77- 83. (in Russian).

[23] N. I. Chernov, "The effectiveness of the apparatus of linear spaces in the logic synthesis of digital structures," Proceedings of the international scientific and technical conferences "Intelligent systems (IEEE AIS'05)" and "Intelligent CAD (CAD-2005",vol.1, pp. 420 –424. (in Russian).

[24] N. I. Chernov and V. Ya. Yugai, "Nonclassical Synthesis of Digital Structures by Tools of Analogous Circuits Engineering," Problems of Today's Analogous Circuits Engineering: The Collected Articles of IX International Scientific-Practical Seminar edited by N.N. Prokopenko, Shakhty, Rostov-on-Don Region: FSBEU HPE "SRSUES" Publishers, 2012, pp. 138 – 143. (in Russian).

[25] N. N. Prokopenko, N. I. Chernov, V. Ya. Yugai, "Base Concept of Linear Synthesis multi-Valued Digital Structures within Linear Spaces," Proceedings of The IS&IT13 Congress, The Scientific Edition in four volumes, Moscow: PhisMathLit publ., 2013, v. 1, pp. 284−289. (in Russian).

[26] N. N. Prokopenko, N. V. Butyrlagin, N. I. Chernov, V. Ya. Yugay, "Synthesis of binary triggers in the apparatus of linear algebra," Izvestiya SFedU. Technical Sciences. No. 2. 2015, pp. 115-125. (in Russian).

[27] N. N. Prokopenko, N. I. Chernov, V. Ya. Yugai, N. V. Butyrlagin, "Linear synthesis of k-valued digital components of the base current logical signals: the principle of generalization", Problems of development of perspective micro and nanoelectronic systems - 2016. Proceedings / under the General editorship of academician RAS A.L. Stempkovskogo. Moscow: IPPM RAS, 2016. (in Russian).

[28] N. I. Chernov, V. Y. Yugai, N. N. Prokopenko, N. V. Butyrlagin, "Basic concept of linear synthesis of multivalued digital structures in linear spaces," Proceedings of IEEE East-West Design and Test Symposium, EWDTS 2014, art. no. 7027045. DOI: 10.1109/EWDTS.2014.7027045

[29] N. N. Prokopenko, N. V. Butyrlagin, N. I. Chernov, V. Ya. Yugai, "The linear concept of logic synthesis of digital IP-modules of control and communication systems," 2015 International Siberian Conference on Control and Communications, SIBCON 2015 - Proceedings, art. no. 7147182. DOI: 10.1109/SIBCON.2015.7147182.

[30] N. N. Prokopenko, N. V. Butyrlagin, N. I. Chernov, V. Ya. Yugai, "Basic linear elements of k-Valued digital structures," 2016 International Conference on Signals and Electronic Systems, ICSES 2016 - Proceedings, pp. 7-12. DOI: 10.1109/ICSES.2016.7847763.

[31] N. N. Prokopenko, N. I. Chernov, V. Yugai, N. V. Butyrlagin, "The element base of the multivalued threshold logic for the automation and control digital devices," Proceedings of 2017 IEEE International Siberian Conference on Control and Communications, SIBCON 2017, art. no. 7998508. DOI: 10.1109/SIBCON.2017.7998508.

[32] N. N. Prokopenko, N. I. Chernov, V. Yugai, N. V. Butyrlagin, "The multifunctional current logic element for digital computing devices, operating on the principles of linear (not boolean) algebra," Proceedings of 2016 IEEE East-West Design and Test Symposium, EWDTS 2016, art. no. 7807723. DOI: 10.1109/EWDTS.2016.7807723.

[33] V. I. Nechaev, "Numeric system," Moscow: Education publ.,1975, 199 p. (in Russian).

[34] V. A. Gorbatov, "Foundations of discrete mathematics," Moscow: Higher school publ., 1986, 311 p. (in Russian).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 191-197 (2018)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Effective Thermal Analysis of Using Peltier Module for Desalination Process

Hayder Al-Madhhachi [*, 1, 2]

[1]*School of Engineering, Cardiff University, CF24 3AA, UK*

[2]*Faculty of Engineering, University of Kufa, Najaf, Iraq*

ARTICLE INFO

Article history:
Received: 05 November, 2017
Accepted: 20 December, 2017
Online: 30 January, 2018

Keywords :
Thermal Analysis
Thermoelectric
Peltier Module
Desalination
Distillation

ABSTRACT

The key objective of this study is to analyse the heat transfer processes involved in the evaporation and condensation of water in a water distillation system employing a thermoelectric module. This analysis can help to increase the water production and to enhance the system performance. For the analysis, a water distillation unit prototype integrated with a thermoelectric module was designed and fabricated. A theoretical model is developed to study the effect of the heat added, transferred and removed, in forced convection and laminar flow, during the evaporation and condensation processes. The thermoelectric module is used to convert electricity into heat under Peltier effect and control precisely the absorbed and released heat at the cold and hot sides of the module, respectively. Temperatures of water, vapour, condenser, cold and hot sides of the thermoelectric module and water production have been measured experimentally under steady state operation. The theoretical and experimental water production were found to be in agreement. The amount of heat that needs to be evaporated from water-vapour interface and transferred through the condenser surface to the thermoelectric module is crucial for the design and optimization of distillation systems.

1. Introduction

Salt water is about 97% of the world's water, but less than 3% is fresh water. In this small fraction, only 0.3% of the fresh water is used by humans. Many places in the world, for example the developing countries, are facing water crises because of population growth and climate change [1-3]. Generating drinkable water from salt water (desalination) is one of the most important approaches to solve this issue without any serious impact on the environment [4-7]. Desalination can be achieved using two methods which are reverse osmosis technology and thermal distillation through evaporation and condensation processes [8-10].

Evaporation and condensation are fundamentally convection heat transfer processes involving phase change. When the temperature of water at given pressure is increased to the saturation temperature, water will evaporate. Similarly, condensation occurs when the temperature of a vapour is reduced below its saturation temperature. In the thermal distillation systems, control the evaporation and condensation rates lead to enhance the water production by increasing the water and vapour temperatures and decrease the condenser temperature. However, these processes consume energy and needs to enhance its performance [11-15].

There are various integrated technologies for water distillation systems, one of these is thermoelectric (TE) technology. Thermoelectric modules have no moving parts and a long life. They are noiseless, easy to control, and compact in size, consume a small amount of energy and so cause less pollution. The principle of the thermoelectric module is based on the Peltier, Seebeck and Thomson effects. The thermoelectric modules can be employed for refrigeration (Peltier), power generation (Seebeck) and temperature sensing (Thomson) [16-20]. The basic concept of thermoelectric modules is to convert thermal energy directly into electrical energy and vice versa. As a temperature gradient is established within a thermoelectric module, voltage difference is produced (under Seebeck mode) and a DC current will be created. As a DC current is applied within a thermoelectric module, temperature difference is established between the hot and cold sides of the module (under Peltier mode). The Peltier modules are

[*]Corresponding Author: Hayder Al-Madhhachi, School of Engineering, Cardiff University, CF24 3AA, UK, +447424082146, Al-MadhhachiHS@cardiff.ac.uk

employed in the thermal distillation systems to generate freshwater by using the cold side of the module to enhance water condensation. Recently, considerable amount of researches have been used thermoelectric technology to increase the water production rate and consequently enhance the system performance [21-28]. Hence, previous thermoelectric distillation systems are not comprehensive enough to properly analyse the evaporation and condensation processes under steady state operation.

This paper considers applying energy conservation to the system including the amount of thermal energy entering and leaving the system under steady state operation. This study aims to recognise the beneficial heat for water production and also governs the evaporation and condensation rates to become equilibrium under controlled conditions.

2. System Description

A designed water distillation system consists of an aluminium chamber placed a top water bath and one thermoelectric module placed between outer surface of the chamber and a water heat exchanger. The cold side of the thermoelectric module is mounted on the top aluminium surface as an inclined condenser, whereas the hot side of the module is attached to the water heat exchanger to enhance heat flow. Figure 1 shows the main parts of the system.

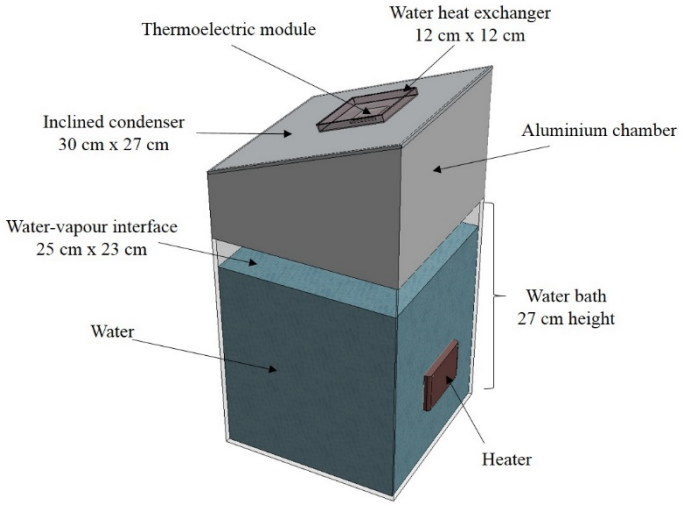


Figure 1. Schematic diagram of the main parts of the water distillation system.

3. Theoretical Model

Theoretical modelling of the water distillation system can be an effective tool for predicting system performance. A set of equations that describes convection heat transfer, fluid flow and phase change are employed to calculate the amount of heat involved in the evaporation and condensation processes and the rates of the evaporation and condensation in the system.

3.1. Assumptions

In the theoretical model, the following assumptions were formed: 1) The thermal properties of water at a pure state; 2) Vapour displays ideal gas behaviour; 3) Water and vapour flow are Laminar; 4) Thermal conductivity of the aluminium condenser is constant. 5) Convection heat transfer coefficients of water and vapour inside the chamber are constant.

3.2. Heat Balance

The model uses energy equations to predict heat added for the evaporation process and that removed by the condensation process from water and vapour phases, respectively as shown in Figure 2a.

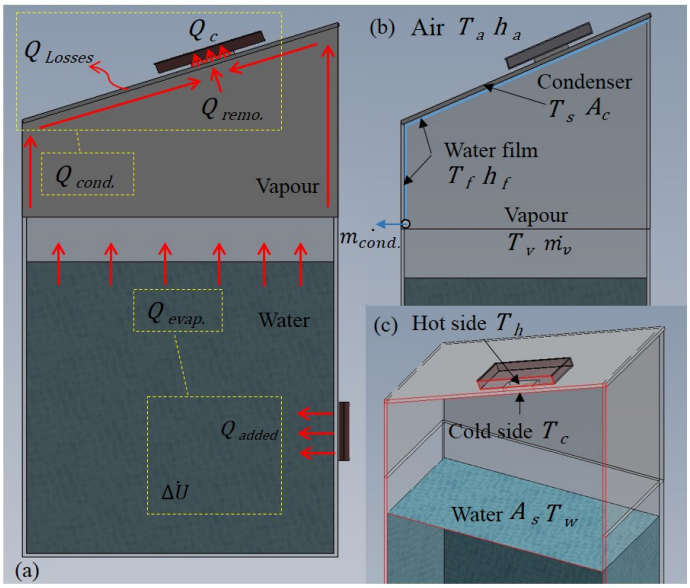


Figure 2. Schematic diagram of: a) The heat balance of the system, b) Evaporation and condensation parameters and c) Hot and cold sides of the thermoelectric module in the system.

Firstly, the conservation of energy at the water-vapour surface for steady state operation is applied. Three heat transfer terms are shown in the system; 1) the heat added Q_{add} to increase or maintain the temperature of the water which is corresponding to the electrical power supply to the heater in the system, 2) The rate of internal energy difference ($\dot{\Delta U}$) of the water, 3) the heat required for the evaporation from the water-vapour surface area which is called in this study the evaporation heat $Q_{evap.}$ Applying the energy conservation to the system, the energy balance takes the form:

$$Q_{add} - Q_{evap.} = \dot{\Delta U} \quad (1)$$

At the isothermal process, water temperature remains unchanged. The heat added will be equal to the evaporation heat ($\dot{\Delta U} = 0$). In that analysis, the following equation is formed to determine the evaporation heat:

$$Q_{add} = Q_{evap.} \quad (2)$$

The rate of produced vapour $\dot{m}_v$ can be calculated based on the evaporation heat and the latent heat of evaporation h_{fg} at a given water temperature using Steam Tables [29].

$$\dot{m_v} = Q_{evap.}/h_{fg} \quad (3)$$

Secondly, the thermal resistance network between the fluids (water film and ambient) and the transmission surface (aluminium condenser) is used to calculate a removed heat $Q_{remo.}$ .The removed heat is a heat released when the vapour condenses to form a water film. It is a function of overall convective heat transfer

coefficient h_c, condenser surface area (A_c = 0.081 m^2) and the temperature difference between the water film T_f (It is a very thin layer of condensed water as shown in Figure 2b) and the condenser surface T_s.

$$Q_{remo.} = h_c A_c (T_f - T_s) \quad (4)$$

$$h_c = 1/(\frac{1}{h_f} + \frac{L}{k_{AL}} + \frac{1}{h_a}) \quad (5)$$

where h_f is the water film heat transfer coefficient which is assumed 100 W/m^2 K (the water film at slow velocity and free convection in the system), h_a is the ambient heat transfer coefficient which is assumed 10 W/m^2 K and k_{AL} is the thermal conductivity of the aluminium condenser which is assumed 205 W/m. K [30]. The temperature of the water film is calculated as an average between the vapour T_v and the condenser surface temperatures [31].

$$T_f = (T_v + T_s)/2 \quad (6)$$

Thirdly, heat flow through the thermoelectric module is determined by the principles of thermoelectrics under Peltier mode [32]. The absorbed heat Q_c at the cold side of the thermoelectric module is:

$$Q_c = \alpha I T_c - K(T_h - T_c) - 0.5RI^2 \quad (7)$$

where α is the Seebeck coefficient of thermoelectric module (assumed to be 0.02 V/K and remains unchanged), I is a DC electric current supply to the thermoelectric module, $(T_h - T_c)$ is the temperature difference between the hot and cold sides of the module (see Figure 2c), K is the thermal conductance of the module (assumed to be 0.1 W/K) and R is the electric resistance of the thermoelectric module (measured value = 0.152 Ω). The first, second and third terms of the right side of Eq. (7) describe the Peltier heat at the cold side of the module, Fourier heat conduction and Joule heating, respectively.

The heat balance between the removed heat, the absorbed heat and losses can be expressed as:

$$Q_{cond.} = Q_{remo.} - Q_c - Q_{losses} \quad (8)$$

where $Q_{cond.}$ is the condensation heat which is the net amount of heat for water condensation and Q_{losses} is the heat losses from the condenser surface to the surrounding. The heat losses is calculated based on the temperature difference between the condenser and the ambient T_a, the convection heat transfer coefficient of the ambient and the area of the condenser surface:

$$Q_{losses} = h_a A_c (T_s - T_a) \quad (9)$$

The rate of the water condensation can then be calculated based on the condensation heat and the latent heat of condensation L_{fg} at the water film temperature using Steam Tables [29].

$$\dot{m_{cond.}} = Q_{cond.}/L_{fg} \quad (10)$$

Finally, it is important to calculate the coefficient of performance of the system *COP*, which is a ratio of the useful cooling provided (condensation heat) to the power required (the total power input to the system P_{total}).

$$COP = Q_{cond.}/P_{total} \quad (11)$$

4. Experimental Setup

A simplified thermoelectric distillation system was designed and constructed as shown in Figure 3a and b, which consists of an Aluminium condenser chamber, a thermoelectric module, a water heat exchanger, a constant temperature water bath and two variable power supplies.

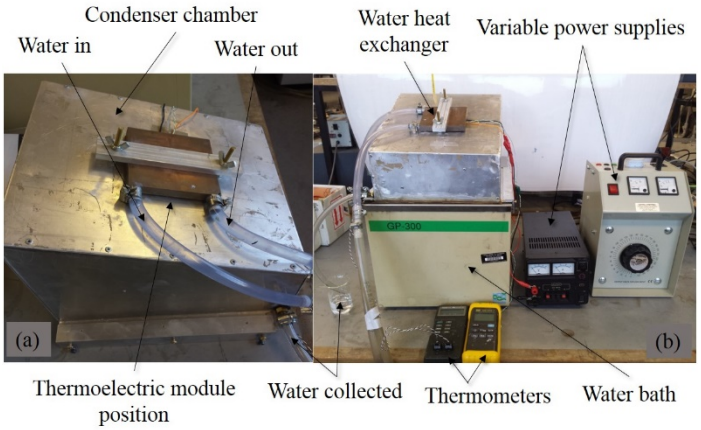


Figure 3. a) Aluminium condenser chamber and b) The experiment setup.

The condenser was fabricated using aluminium sheet (1.5 mm thickness). It was placed on the top of the water bath. One thermoelectric module (4 cm x 4 cm) was used in this study. The water bath and the module were powered by variable power supplies. Cold water was circulated through the water heat exchanger to release the absorbed heat of the module. For effective heat transfer rate, a heat sink compound was used to paste the hot and the cold sides of the module to the water heat exchanger and the condenser, respectively.

Five thermocouples were placed at five different positions in the system to measure the temperatures of the air, vapour, condenser surface and hot and cold sides. The water production was collected and measured every 10 minutes. To examine the experiments more precisely, specifications of the components used in the experiments, together with the module properties and manufacturers information are listed in Table 1. Accuracies, ranges and standard uncertainty of measured instruments are listed in Table 2.

Table 1. Properties of the components used in the experiments.

Component	Properties	Model / Supplier
Thermoelectric Module	Area: 40 x 40 mm^2 Thickness: 3.3 mm Current = 8.5 A and Voltage = 15.4 V	CP-14-127-045 Laird Technologies
Water Bath	Fluid Temperature = -30 – 150 °C, Tank Size = 24 L, Voltage = 115 V and Current = 11.5 A	GP-300 NESLAB

Table 2. Accuracies, ranges and standard uncertainty of the measurement instrumentations.

Instrument	Accuracy	Range	Standard uncertainty
Thermocouple (type K)	0.1 ºC	-75 - 250 ºC	0.06 ºC
Thermometer Fluke 52	0.05 ºC	-200 – 1372 ºC	0.03 ºC
Thermometer RS 206	0.1 ºC	-100 – 1000 ºC	0.05 ºC
Weir 423 power supply	0.01 V 0.002 A	0-15 V 0-2 A	0.006 V 0.001 A
Portable variable transformer CMCTV10	0.01 V 0.02 A	0-300 V 0-10 A	0.005 V 0.01 A

5. Results and Analyses

The experiments have been conducted at constant water flow rate through the water heat exchanger. The thermoelectric module was powered at constant voltage of 10 V and current of 2 A. In all experiments, the water production from the system is equalized the water condensation. All experiments were repeated three times for accuracy under steady state operation.

5.1. Water Production at Different Water Temperatures

Figure 4 shows the water production at different water temperatures under one-hour system operation, with and without using thermoelectric module for cooling process (without using TE module means using a cold water in the water heat exchanger for cooling process). The water production was increased when the water temperatures increased from 40 ºC to 70 ºC. It can be seen also from the Figure that the water production was enhanced twice when using the thermoelectric module for cooling.

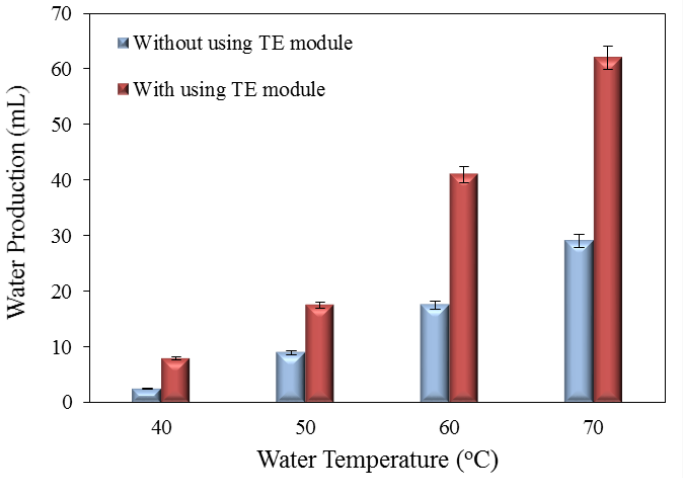


Figure 4. Water production under one-hour system operation at different water temperatures.

5.2. Distillation System at Constant Water Temperature

These experiments have been conducted at constant water temperature 50 ºC. This temperature was chosen to study as an average of the maximum range of the water temperatures in solar stills due to the solar irradiation in Middle East in summer (45 ºC – 55 ºC) [33-36]. For steady state operation, the water bath was powered at constant voltage of 40 V and current of 0.25 A in order to keep water temperature at 50 ºC. The thermoelectric module was also powered at constant voltage of 10 V and current of 2 A in order to keep the cold side temperature of the module at 28 ºC. Therefore, the evaporation heat, the absorbed heat and the total input power to the system are constant.

5.2.1 Water Production Reliability

Figure 5 shows the water production after three hours of the thermoelectric system operation. It can be seen from the figure that the water production was approximately constant with time which confirms the condensation heat was also constant.

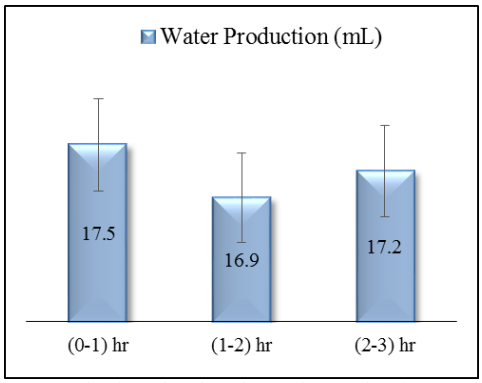


Figure 5. Water production after three hours thermoelectric system operation.

5.2.2 Thermal Behaviour of the System Components

Five different temperatures were measured in the water distillation system during the first hour of thermoelectric operation. The temperatures including; the condenser surface, vapour, ambient, hot and cold sides of the thermoelectric module. It was noticed that all temperatures except the ambient have same behaviour with time as shown in Figure 6.

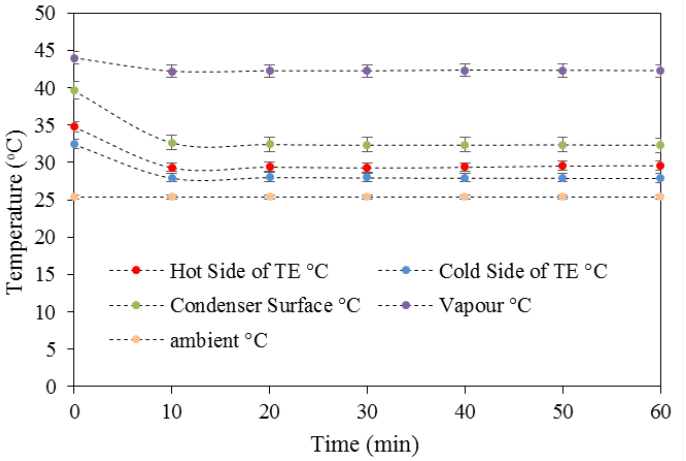


Figure 6. Temperatures variation of the system components.

It was found that the cold side temperature was lower than the condenser surface temperature which indicates that the thermoelectric module was provided cooling to the condenser. The temperature difference between the hot and cold sides of the thermoelectric module was small which means that the thermoelectric module pumped properly the heat from the condenser chamber to the water heat exchanger. At the beginning of the experiments, there was a decrease of the temperatures of

the condenser, the hot and cold sides because of the response of the thermoelectric operation. After 10 minutes of the thermoelectric system operation, the temperatures were remained approximately constant.

5.2.3 Validation with the Experimental Data

Figure 7 shows the theoretical and experimental amount of water production during the first hour of thermoelectric system operation. The water condensation rate was calculated theoretically using equation (10) at constant latent heat of condensation (at 37.3 ºC water film temperature). The water production rate (condensed and collected in the system) was measured experimentally every 10 minutes. The theoretical model predicts well the rate of water production with reasonable agreement. There is a slight difference between the experimental and calculated theoretical values and this is due to losses during the evaporation and condensation process.

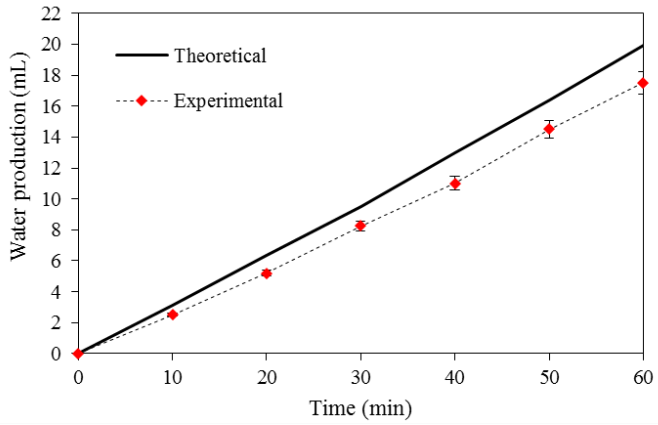


Figure 7. Water production validation with the experimental data.

5.2.4 Percentage of the Beneficial Heat for Water Production

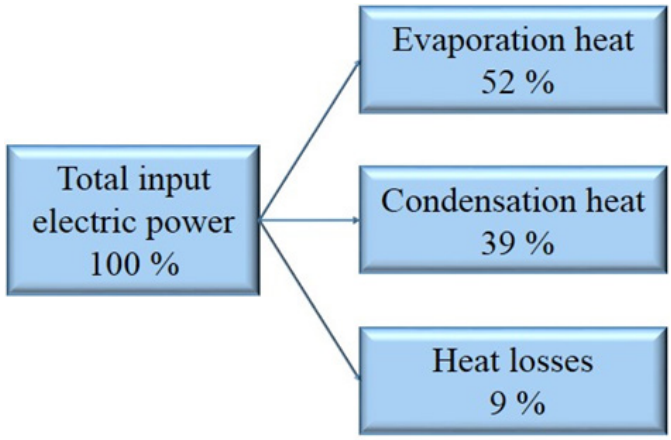


Figure 8. The percentage of the beneficial heat in the system.

Figure 8 shows the heat analysis in the water distillation system under steady state operation. It can be seen that the percentage of the beneficial heat, which is used for the water production, was about 91% based on equation (2) and equation (8), while the percentage of the heat losses to the surroundings was about 9% based on equation (9). The coefficient of performance of the system was 0.3 based on equation (11) under one-hour thermoelectric system operation.

5.2.5 Equilibrium between Evaporation and Condensation Rates

The rate of evaporation is dependent on the surface area of the water-vapour interface and the water temperature. When these factors remain constant, the rate of evaporation will be constant. As well as, the rate of condensation is dependent on the surface area of the condenser and the temperature of condenser (which is entirely dependent on the temperature of the cold side of the thermoelectric module). When these three factors are constant, the rate of condensation will be constant.

Figure 9 shows the evaporation and condensation rates for one-hour system operation at different cold side temperature based on equation (3) and equation (10), respectively. The evaporation rate was maintained steady at constant heat added (10W) and water temperature (50 ºC). When the cold side temperature of the module was decreased from 28 ºC to 22 ºC by increased the current supply to the module from 2 A to 3 A, there was a continual increase in the condensation rate.

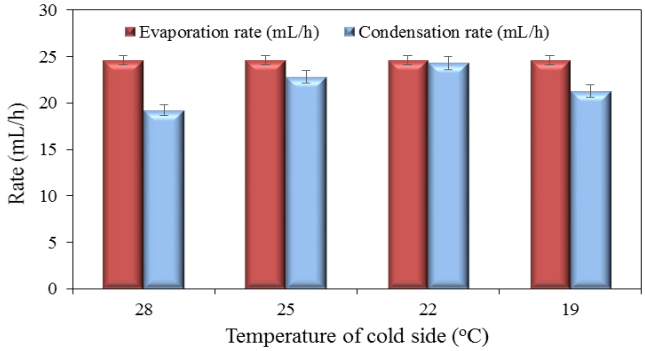


Figure 9. Equilibrium between evaporation and condensation rates.

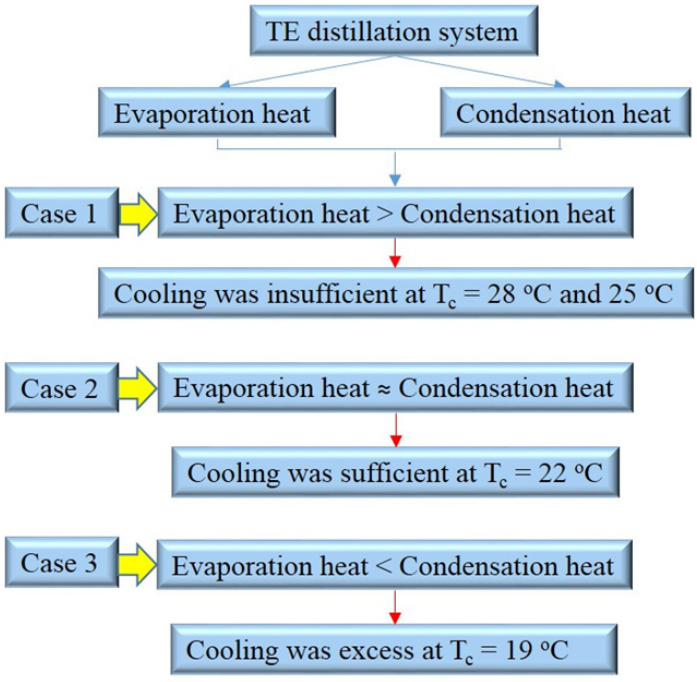


Figure 10. Heat transfer analyses of the system.

The rates of the evaporation and condensation became approximately equilibrium at 22 ºC cold side temperature and 50

ºC water temperature. When the current supply increased to 3.5 A (19 ºC cold side temperature), the condensation rate was decreased with increasing the absorbed heat at the cold side of the thermoelectric module. However, in this case, the cooling was excess.

Based on the results in Figure 9, the heat transfer analyses of the water distillation system is shown in Figure 10. As a conclusion, these results show that the heat involved in the evaporation and condensation processes are an important consideration in the design of an effective distillation system.

6. Conclusions

This study has been investigated to apply the law of conservation of energy to a water distillation system using a thermoelectric module. A theoretical model has been developed to predict the rates of the evaporation and condensation. The theoretical analysis of the heat required for evaporation and condensation processes in the thermoelectric water distillation system was carried out to assist in design of high performance thermal distillation system. This analysis is an effective tool for energy saving issues. A prototype distillation system integrated with one thermoelectric module was designed and fabricated to measure the rates of the evaporation and condensation under one-hour steady state. The results show that there is a reasonable agreement between the theoretical model and the experimental data, the percentage of the beneficial heat for the water production was about 91% and the rates of the evaporation and condensation became approximately equilibrium at 22 ºC cold side temperature and 50 ºC water temperature.

Acknowledgment

The author wishes to acknowledge financial support from the Ministry of Higher Education and Scientific Research, University of Kufa, Iraq. Also, the author would like to thank Professor. Gao Min, Dr. Jorge García-Cañadas, Dr. Matthew Phillips and Dr. Tanuj Singh and the Thermoelectric and Photovoltaic Group in Cardiff School of Engineering who gave scientific discussions.

References

[1]. Loucks, D.P., 2017. Managing water as a critical component of a changing world. Water Resources Management, pp.1-12.

[2]. Petersen, L., Heynen, M. and Pellicciotti, F., 2017. Freshwater Resources: Past, Present, Future. The International Encyclopedia of Geography.

[3]. Manju, S. and Sagar, N., 2017. Renewable energy integrated desalination: A sustainable solution to overcome future fresh-water scarcity in India. Renewable and Sustainable Energy Reviews, 73, pp.594-609.

[4]. Sommariva, C., 2017. State of the Art and Future Applications of Desalination Technologies in the Middle East. In Water, Energy & Food Sustainability in the Middle East (pp. 107-124). Springer International Publishing.

[5]. Chandrashekara, M. and Yadav, A., 2017. Water desalination system using solar heat: A review. Renewable and Sustainable Energy Reviews, 67, pp.1308-1330.

[6]. Ghaffour, N., Bundschuh, J., Mahmoudi, H. and Goosen, M.F., 2015. Renewable energy-driven desalination technologies: A comprehensive review on challenges and potential applications of integrated systems. Desalination, 356, pp.94-114.

[7]. Sharon, H. and Reddy, K.S., 2015. A review of solar energy driven desalination technologies. Renewable and Sustainable Energy Reviews, 41, pp.1080-1118.

[8]. Likhachev, D.S. and Li, F.C., 2013. Large-scale water desalination methods: a review and new perspectives. Desalination and Water Treatment, 51(13-15), pp.2836-2849.

[9]. Drioli, E., Ali, A. and Macedonio, F., 2015. Membrane distillation: Recent developments and perspectives. Desalination, 356, pp.56-84.

[10]. Shahzad, M.W., Burhan, M. and Ng, K.C., 2017. Pushing desalination recovery to the maximum limit: Membrane and thermal processes integration. Desalination, 416, pp.54-64.

[11]. Al-Sulttani, A.O., Ahsan, A., Rahman, A., Daud, N.N. and Idrus, S., 2017. Heat transfer coefficients and yield analysis of a double-slope solar still hybrid with rubber scrapers: An experimental and theoretical study. Desalination, 407, pp.61-74.

[12]. Eldeeb, R., Aute, V. and Radermacher, R., 2016. A survey of correlations for heat transfer and pressure drop for evaporation and condensation in plate heat exchangers. International Journal of Refrigeration, 65, pp.12-26.

[13]. Sathyamurthy, R., El-Agouz, S.A. and Dharmaraj, V., 2015. Experimental analysis of a portable solar still with evaporation and condensation chambers. Desalination, 367, pp.180-185.

[14]. [14] H. Al-Madhhachi, M. Phillips and G. Min, "Validation of Vapour/Water Production in a Thermoelectric Distillation System" in Proceedings of the 3rd World Congress on Mechanical, Chemical, and Material Engineering (MCM'17), Roma, Italy, 2017. ISSN: 2369-8136. DOI: 10.11159/htff17.117.

[15]. Mishra, D.R., Tiwari, A.K. and Sodha, M.S., 2016. Mathematical modeling and evaluation of new long single slope still for utilization of hot wastewater. Applied Thermal Engineering, 108, pp.353-357.

[16]. Elsheikh, M.H., Shnawah, D.A., Sabri, M.F.M., Said, S.B.M., Hassan, M.H., Bashir, M.B.A. and Mohamad, M., 2014. A review on thermoelectric renewable energy: Principle parameters that affect their performance. Renewable and Sustainable Energy Reviews, 30, pp.337-355.

[17]. Chein, R. and Huang, G., 2004. Thermoelectric cooler application in electronic cooling. Applied Thermal Engineering, 24(14), pp.2207-2217.

[18]. Bell, L.E., 2008. Cooling, heating, generating power, and recovering waste heat with thermoelectric systems. Science, 321(5895), pp.1457-1461.

[19]. Goldsmid, H.J., 2016. Theory of Thermoelectric Refrigeration and Generation. In Introduction to Thermoelectricity (pp. 9-24). Springer Berlin Heidelberg.

[20]. Riffat, S.B. and Ma, X., 2003. Thermoelectrics: a review of present and potential applications. Applied thermal engineering, 23(8), pp.913-935.

[21]. Al-Madhhachi, H. and Min, G., 2017. Effective use of thermal energy at both hot and cold side of thermoelectric module for developing efficient thermoelectric water distillation system. Energy Conversion and Management, 133, pp.14-19.

[22]. Moh'd A, A.N. and Al-Ammari, W.A., 2017. Utilizing the evaporative cooling to enhance the performance of a solar TEG system and to produce distilled water. Solar Energy, 146, pp.209-220.

[23]. Demir, M.E. and Dincer, I., 2017. Development of an integrated hybrid solar thermal power system with thermoelectric generator for desalination and power production. Desalination, 404, pp.59-71.

[24]. Rahbar, N., Esfahani, J.A. and Asadi, A., 2016. An experimental investigation on productivity and performance of a new improved design portable asymmetrical solar still utilizing thermoelectric modules. Energy Conversion and Management, 118, pp.55-62.

[25]. H. S. S. Al-Madhhachi, "Solar Powered Thermoelectric Distillation System," Ph.D Thesis, Cardiff University, 2017.

[26]. Date, A., Gauci, L., Chan, R. and Date, A., 2015. Performance review of a novel combined thermoelectric power generation and water desalination system. Renewable energy, 83, pp.256-269.

[27]. Aberuee, M.J., Baniasadi, E. and Ziaei-Rad, M., 2017. Performance analysis of an integrated solar based thermo-electric and desalination system. Applied Thermal Engineering, 110, pp.399-411.

[28]. H. Al-Madhhachi, M. Prest and G. Min, G., " Evaluation of the convection heat transfer coefficient in a thermoelectric distillation system" in 2016 International Conference for Students on Applied Engineering (ICSAE), Newcastle UK, 2016, (pp. 213-217). IEEE. DOI: 10.1109/ICSAE.2016.7810190.

[29]. Keenan, J.H. and Keenan, J.H., 1992. Steam tables: thermodynamic properties of water, including vapor, liquid, and solid phases (English units). United States of America: Second edition. John Wiley & Sons.

[30]. Bejan, A., 2013. Convection heat transfer. John wiley & sons.

[31]. Bergman, T.L. and Incropera, F.P., 2011. Fundamentals of heat and mass transfer. John Wiley & Sons.
[32]. Rowe, D.M. ed., 1995. CRC handbook of thermoelectrics. CRC press.
[33]. Phadatare, M.K. and Verma, S.K., 2007. Influence of water depth on internal heat and mass transfer in a plastic solar still. Desalination, 217(1-3), pp.267-275.
[34]. El-Sebaii, A.A., Al-Ghamdi, A.A., Al-Hazmi, F.S. and Faidah, A.S., 2009. Thermal performance of a single basin solar still with PCM as a storage medium. Applied Energy, 86(7), pp.1187-1195.
[35]. Rahbar, N. and Esfahani, J.A., 2012. Experimental study of a novel portable solar still by utilizing the heatpipe and thermoelectric module. Desalination, 284, pp.55-61.
[36]. Baalbaki, A., Ayoub, G.M., Al-Hindi, M. and Ghauch, A., 2017. The fate of selected pharmaceuticals in solar stills: Transfer, thermal degradation or photolysis. Science of The Total Environment, 574, pp.583-593.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 198-204 (2018)

www.astesj.com

ASTESJ

ISSN: 2415-6698

Estimation of digital protection devices applicability on basis of multiple characterizing parameters

Dimitar Bogdanov*

Technical University of Sofia (TUS), Faculty of electrical engineering, Chair "Electrical power engineering" "Relay Protection" Laboratory

A R T I C L E I N F O

Article history:
Received: 17 November, 2017
Accepted: 10 January, 2017
Online: 30 January, 2018

Keywords :
Electrical power system
Intelligent electronic device (IED), protection devices,
Selection, raking, criteria

A B S T R A C T

The contemporary electrical power systems (EPS) impose increased requirements for the functionality of the protection systems. The necessity of improved EPS stability is in some extent resulting of the increased integration of renewable sources of electrical energy. The future grid development gives perspective for connection of more converter based generations. The power electronic schemes and associated functional requirements impose necessity of high speed, sensitive, selective and reliable operation of the protection devices. These requirements have always been target of the protection equipment producers and grid operators. The electronic converting schemes specifics impose these requirements for the protection devices in more straightened way, as the converter connected generator may need to trip in shorter time than classical machine generator. In the article is presented a generalized overview of some of the characteristics of the digital "relay" protection devices, and approach for device selection is proposed. Investment planning may utilize such approach in order to have an optimal design from financial point of view.

1 Introduction

During the last decades the protection systems went through extensive technological advance. Practically the term "relay protection" becomes a bit obsolete, or depicting only the particular sector of automation of the EPS. There is a sustainable tendency more converter based generations to be grid connected [1]. Several cases of severe black-outs [2,3,4,5] imposed increased requirements in respect to EPS protection. The results of post-fault analysis for such events indicated, that normative documents in many countries were in some extent obsolete. In result many new rules and design principles were adopted to enhance the EPS stability, power quality improvement of operation and maintenance flexibility [6,7]. Cyber security is also an important issue for the contemporary EPS automation schemes, and in result of the tendency for transfer to digital technologies, this problem will stand for the future as well [9,10,11]. The electrical/electronic equipment vulnerability to other natural phenomena like extreme solar activity [10] also imposes challenges for the protection systems. "Magnetic storms/solar protuberances" have already impacted the operation of EPS protection and control systems [10]. The protection devices recordings for fault events can be a valuable source of data for the analysis of the pre-history and consequences of the event [8,9,12].

The functionality of the "relay" protection systems can be observed from different aspects. The optimal design solution for particular scheme of protection system can be estimated by different indices as functional applicability, reliability, compatibility with site conditions, maintenance, MMI (Man-Machine Interface) specifics, etc.

In order to achieve an optimal selection of protection devices / IEDs for particular application, object of interest is the characterizing criteria to be systematized and evaluated with "structured" approach. First of all such approach can help to create a "template form" for estimation equipment compliance with user's requirements. The elaboration of such approach may also ease the selection of equipment out of several alternatives. If preliminary set of requirements / criteria for protection equipment selection (or protection and control) is available, it can be systematized in groups with the respective listed characteristics.

The paper is based on study of 5 different vendors of digital protection devices / IEDs. All studied devices were tested for

*Corresponding Author: Dimitar Bogdanov, Email: dbogdanov@tu-sofia.bg

functional behavior: fault conditions were simulated, connected to PC with firmware for data transfer of settings, logs of events analyzed, etc. The samples of equipment were analyzed in laboratory conditions.

Each of the protection device characteristics can be assigned to a more generalized group, and in such manner they can be presented as:

1) Operational principle, functional applicability, applicability for the particular object (scheme);
2) Compliance with the respective standards, state /grid operator, utility/ requirements, and client's specific rules;
3) Reliability indices;
4) Self-test function, level of self-test, self-test signaling;
5) Overall construction – general characteristics;
6) Overall construction – internal structure;
7) Overall construction – set of functions (electrical, technological protections);
8) Intelligent Electronic Devices (IED) characteristics;
9) Installation location conditions compatibility;
10) Compatibility with power supply;
11) Compatibility with peripheral devices;
12) Firmware (software);
13) Man – machine interface (MMI) characteristics;
14) Interfaces to other systems;
15) Service time, without necessity of maintenance / servicing;
16) Data / event registration; "oscillographic" recorder functions /wave form records/;
17) Functionality to operate in centralized/decentralized systems for protection and control;
18) Functionality to receive (and redistribute) centralized synchronization;
19) Functionality to provide functions for Wide Area Protection (WAP), grid (EPS) level protection functions (load shedding, etc.);
20) Functionality for control of commutation apparatus, interlocking between devices, etc.;
21) Modular design, options for upgrade and extension;
22) Service life, guaranteed under the specified operating conditions;
23) Guaranteed period for service, procurement of spare parts, firmware update / support;
24) Unauthorized access protection;
25) References for application, other;
26) Financial aspects.

2 Proposed method for estimation of digital protection devices applicability

In result of study on the vendors proposed devices, the required characteristics and protection devices specifics by users / regulations, generalization was made. The above listed general characteristics (1-26) do not cover all possible categories to estimate the devices specifics. They are presented as general list of IED functionality features, but not pretending to envelope in completeness all applicable characteristics. The qualitative estimation of the level of fulfillment of the designer / investor requirements for each characteristic can be ranked as:

$$e_i = \{0 \div r_i\}, \quad (1)$$

where r_i is the maximal ranking for each characteristic. For the presented example list of 26 groups of generalized characteristics, for each group the value E_c can be calculated. It can be calculated on basis of the sum of score ranking points e_i for each sub-characteristic pertaining to the respective group:

$$E_c = \sum_{i=1}^{n} e_i, \quad (2)$$

where n is the number of grouped "sub-characteristics".

For equipment evaluation based on 26 generalized characterizing features, can be written:

$$E_S = \sum_{j=1}^{m} E_{cj}, \quad (3)$$

Where m for the presented set of characteristic groups is 26.

In Table 1 is presented the set of characteristic groups, detailed in sub-characteristics. In the presented proposal for protection equipment estimation 26 groups are included, but definitely this number can be extended or limited in respect to the needed range. In the right column are presented the maximal ranking points, giving options to estimate the 100% criteria fulfillment. For "fulfillment" is assumed the operator's requirement, accomplishment to achieve the needed protection device functionality.

Table 1 IED set of characteristic groups

N	**Characteristic (group of characterizing parameters) / Sub-characteristics**	**Ec/r$_i$**
1	Operational principle, functional applicability, applicability by particular object (scheme)	**100**
	• Correspondence of the set of functions to the protected object	25
	• Completeness of the set of functions in respect to the necessary full set	25
	• Applicability of the specific operational principle to the object. Amicability for the range of variation of the operating values.	25
	• Applicability of the device to the particular technical parameters of the interfacing connections	25
2	Compliance with the respective standards and regulations, state requirements /EPS operator (TSO), utility/ and client's specific rules	**100**
	• Compliance with applicable international standards and regulations	50
	• Compliance with applicable national / client norms, rules and standards	50
3	Reliability indices	**200**
	• MTTF (Mean Time To Failure)	25
	• MTBF (Mean Time Between Failures)	25
	• MTTR (Mean Time To Repair)	25

N	Characteristic (group of characterizing parameters) / Sub-characteristics	Ec/r_i
	• MTTDF (Mean Time To Detect Failure) /related also to self-test, annunciation/	25
	• POFTT (Probability of Fail to Trip) /can be calculated on probabilistic basis, as internally blocked by self-test relay may fail to clear fault/	100*
4	Self-test function, level of self-test, self-test signaling	**100**
	• General hardware self-test	30
	• Data processing self-test	25
	• Peripheral analog signal circuits self-test / diagnostics	25
	• External circuit supervision (trip, measurement circuit, etc.) / diagnostics	20
5	Overall construction – general characteristics	**120**
	• Compatibility of the hardware construction to intended place of installation	35
	• Degree of protection (IP) of casing / peripheral units / terminal strips applicability	20
	• Stability of operating parameters in respect to ambient conditions ranges	20
	• Electromagnetic compatibility	20
	• Seismic qualification	15
	• Thermal requirements / necessity of forced ventilation / cooling, preheating etc.	10
6	Overall construction – internal structure	**140**
	• Internal structure organization. Processing boards (CPUs) number. Task managing for calculation / operational memory optimal resource utilization.	20
	• Internal bus - type, parameters, expandability.	20
	• Available type of peripheral boards, parameters, interchangeability	10
	• Processor(s) type, speed, productivity	10
	• Bit range of the data bus	10
	• Operational memory (RAM), type, installed capacity, maximal size	10
	• Absence of rotating parts in forced cooling for CPU/other ICs	10
	• ADC resolution	10
	• Signal processor /type, characteristics/	10
	• Sampling rate of the measured quantities	10
	• Non-volatile support for memory / data back-up in case of power interruption	10
	• EEPROM (flash memory) for data recording / SSD	10
7	Overall construction – set of functions (electrical, technological protections)	**20**
	• Options for adding (activation) of additional functions	10
	• Back-up functions	10
8	Intelligent Electronic Devices characteristics	**100**
	• IED class device functionality	100
9	Installation location conditions compatibility	**100**
	• Operability in the ambient conditions: EMC, altitude, humidity, temp. range, etc.	100
10	Compatibility with power supply	**60**
	• Power supply ratings (voltage, type of voltage, range)	20
	• Power consumption	20
	• Internal power supply redundancy	20
11	Compatibility with peripheral devices	**100**
	• Measurement (instrumental) transformers cores load	20
	• Availability and compatibility of the analog I/Os	20
	• Availability and compatibility of the discrete I/Os	20
	• Ratings of the input parameters values, options for software adjustment	10
	• Analog inputs range and compatibility	10
	• Permissible tolerances of the measurement (instrumental) tr-s	10
	• Options for alternative sensors detachment (“non-classical” voltage/current, etc. measurement)	10
12	Firmware (software)	**100**
	• Software for IED configuration, option(s) for preliminary configuration of all settings (“online” and “off-line”), configuration settings in a file, to be uploaded in IED(s).	30
	• Graphical presentation (visualization) of the characteristics of the protection functions adjustment	20
	• Functionality for recordings of events, “oscillographic” records, time stamping.	20
	• Options for upgrade, expanding and compatibility with other versions.	10
	• Compatibility of the software with OSs for PCs.	10
	• Option to operate with the particular language / alphabet.	10
13	Man – machine interface (MMI) characteristics	**60**
	• Display for visualization – type, characteristics (LCD, TFT, LED, …) (backlight)	10
	• Signaling LEDs – number, configuration options. Self-test status LED.	10
	• Control buttons (number, ergonomic, durability)	10
	• Interfaces for the operator (Ethernet, USB, RS232, RS485)	10
	• Protection device operation status indicator (in service / internal fault, etc.)	10
	• Option to operate with the particular language / alphabet	10
14	Interfaces to other systems	**60**
	• Ethernet	20

N	Characteristic (group of characterizing parameters) / Sub-characteristics	Ec/r_i
	• Optical	20
	• RS232, RS485	10
	• Other	10
15	Service time, without necessity of maintenance / servicing	**50**
	• Not necessary to perform periodic maintenance (or for particular time interval)	20
	• Not necessary to perform periodic test of adjustments (or for particular time interval)	10
	• Not necessary to perform periodic cleaning of some components /or for particular time interval/	10
	• Period of service, after which some components have to be changed – internal battery, surge arresters, filter capacitors in power supply unit, air filters, fans, etc.	10
16	Data / event registration; "oscillographic" recorder functions	**70**
	• Options for registration (recording) of data	30
	• Number of signals being recorded as separate channels	10
	• Number of records which can be stored	10
	• Records resolution (density of sampling points)	10
	• Options for visualization and analysis of the recorded data, supporting software tools	10
17	Functionality to operate in centralized/decentralized systems for protection and control	**60**
	• IEC 61850 compatibility (IEC 61870, etc.)	60
18	Functionality to receive (and redistribute) centralized synchronization (GPS / IRIG-B, other applicable master synchronization signals)	**30**
	• Functionality to connect to common – site data exchange systems	10
	• Functionality to connect to common – site synchronization system	10
	• Functionality to connect to global synchronization (GPS, other)	10
19	Functionality to provide functions for Wide Area Protection (WAP). EPS (Grid) level protection functions;	**40**
	• Protection functions for local measures to achieve global EPS stability improvement, limitation of the extension of fault processes and unacceptable modes of operation	10
	• Functionality for participation in Wide Area Protection (WAP)	10
	• Functionality for participation in Wide Area Measurement System (WAMS)	10
	• Functionality to receive / generate / transmit signals for anticipated response	10
20	Functionality for control of commutation apparatus, interlocking between devices, etc.	**100**
	• Functionality to replace or duplicate the conventional relay/contact based interlocking schemes with "fail-safe" guaranteed design	100
21	Modular design, options for upgrade and extension	**20**
	• Availability of spare analog / discrete inputs. Optional functionality to integrate signals from technological measurement system (I&C)	10
	• Compatibility between protection devices from different generations (versions). Similarity in the adjustment and maintenance principles.	5
	• Availability of internal bus, which allows adding additional expansion modules	5
22	Service life, guaranteed under the specified operating conditions	**50**
	• Service life, guaranteed by the producer at rated operating conditions, limiting conditions	40
	• Shelf life, storage conditions	10
23	Guaranteed period for service, procurement of spare parts, firmware update / support	**40**
	• Period of provided maintenance	10
	• Period of provided spare parts	10
	• Possibility modules to be replaced /repaired on-site	10
	• Possibility modules to be replaced /repaired in factory	10
24	Unauthorized access protection	**100**
	• Protection against unauthorized access, options for different levels of access, cyber security aspects	100
25	Financial aspects	**30**
	• Cost of equipment	10
	• Relative cost of the protection equipment in respect to protected object	10
	• Maintenance cost	10
26	References for application, other	**50**
	• References for positive operational experience, other non-technical aspects	50
	Total maximal rating points, Es	**2000**

** The reliability indices can be adapted in dependence of the particular application / data available.*

The importance of the presented above protection device characteristics depends on the particular case of implementation. If "weighting coefficients" are used as multipliers for each group of parameters, they can have different relative distribution of the ranking values. In some cases not the devices themselves, but the accessories and the software (firmware) can have a decisive importance. The software is a key feature for the client in some

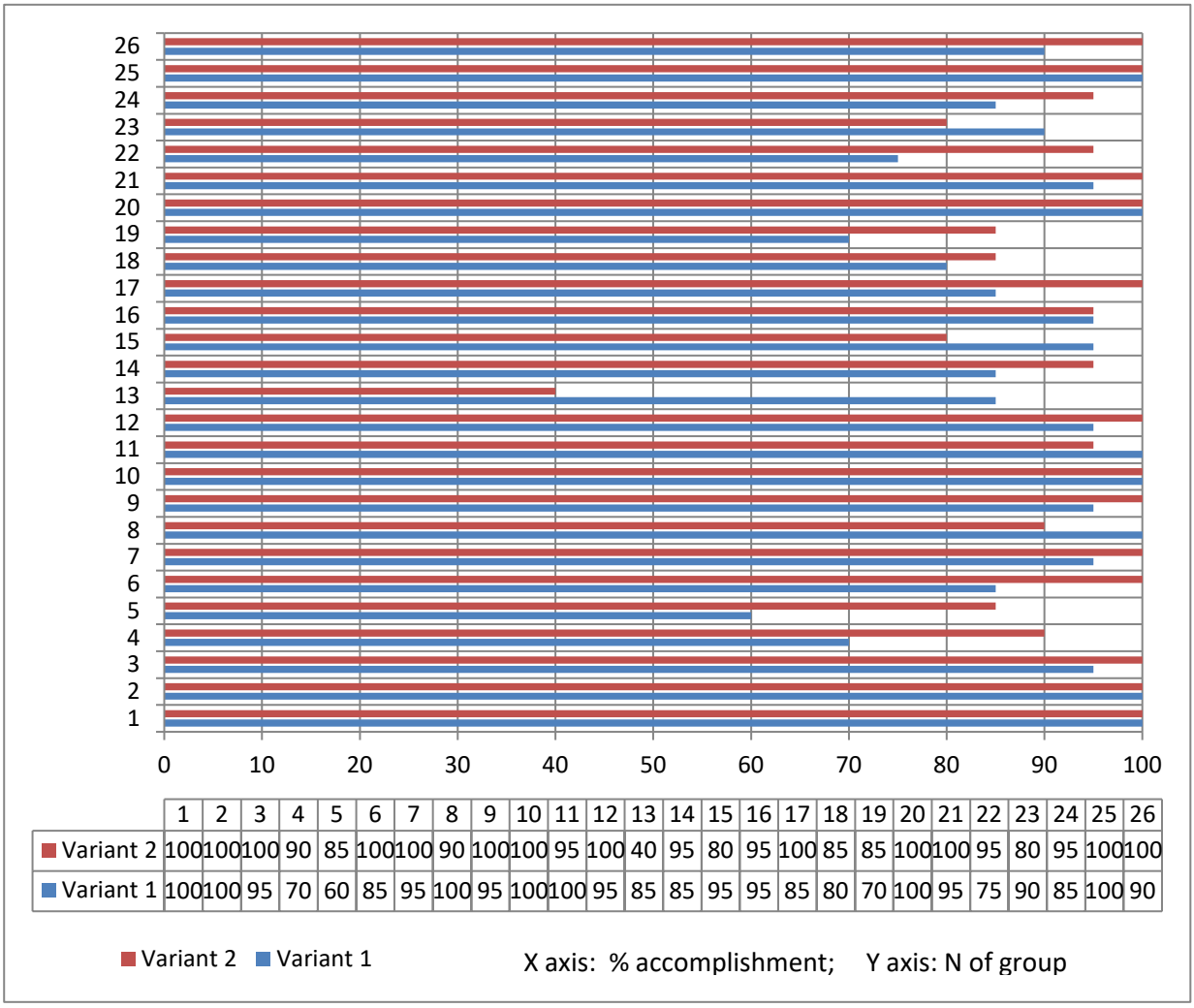
cases. Some companies offer free firmware (for settings, download of records, etc.), some companies offer it as additional option, not free of charge. There are specific occasions, where even a single feature can be a serious obstacle for proper handling of the device: for instance – the device may comply with almost all requirements, but if there is no option for particular language – this could be a critical obstacle for implementation. Some IED producers require payment for the software or part of the software functions. If these expenses are not counted in the "grand total" for protection device / system procurement, the financial comparison of possible options won't be precise. Depending on the scope of functionality, different software require different computing resources. Some firmware products are supplied to support wide spectrum of devices, but the software complexity can cause slow data transfer, inconvenience during work and extended requirements for PC resources. It is relatively common practice the vendor companies to provide basic operating platform, and the respective software modules for the particular type of protection devices (connectivity packages, product specific libraries) shall be additionally installed.

The software itself can provide different approaches for handling of the data, visualization screens, etc. The practice shows, that the software which provides informative and clear graphical presentation of the device settings facilitates the work process and reduces the probability for commissioning specialist / operational personnel mistakes.

3 Theoretical results

The scope of criteria presented in Table 1 gives a total maximal "rating" of 2000 points. The distribution of the ranking points between the respective sub-groups shall be revised for a particular case. The specifics of a particular project may impose some features of the selected devices to be more important in conditions of particular application. The content of the set of criteria may also vary.

In order to achieve realistic results by utilization of the proposed approach, it is important to have in mind not only the "quantity" of ranking points in general, but the balance between the accomplishments of the requirements by each position in general. Figuratively, this means that an option for device configuration can have a higher general score than other, but due



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Variant 2	100	100	100	90	85	100	100	90	100	100	95	100	40	95	80	95	100	85	85	100	100	95	80	95	100	100
Variant 1	100	100	95	70	60	85	95	100	95	100	100	95	85	85	95	95	85	80	70	100	95	75	90	85	100	90

Figure 1. Example for comparison of hypothetical options for protection devices selection.

to the facts that some criteria are maximally covered and other are not covered at all. Additional coefficients can be added in order to estimate the distribution of the values between the groups.

In Fig. 1 is presented an example for "application" of the proposed approach for estimation of digital protection devices applicability on basis of multiple characterizing parameters, based on two variants: "Variant 1" and "Variant 2". The "grand total" score for Variant 1 is 89.42% and for Variant 2 - 92.7%. By absolute values – "Variant 2" shall be preferred. If a "threshold level" of 50% accomplishment is assumed, the screening of the individual groups of characteristics scoring indicates that for selection criteria N 13 - "Variant 2" fails with 40% accomplishment.

The systematized in such manner criteria based on the characteristics of the protection devices (or systems) can be used in two aspects:

- **Estimation for equipment compliance to predefined requirements**. The presence of particular feature of the protection device and/or compliance of the required parameter(s) brings the respective estimation (analysis of compliance) ranking points. After estimation of all of the features is formed a general (total) sum of ranking points (Es) is calculated. The sum of ranking points can be used as overall estimated criteria for the compliance of the respective device (system) with the preliminary defined criteria.
- **Comparison of two (or more) options**. The estimation can be made as every characteristic of the different optional types of equipment is compared to reference criteria. Possible approach is also to proceed with mutual comparison of the optional equipment characteristics. This approach can support the selection of optimal variant for protection devices (systems).

Some characteristics can be detailed further in the aspects of functionality to receive / generate / transmit signals for anticipated response. In case of occurrence of faults or other unacceptable processes, such functionality can help reducing the spread of disturbance in the EPS. Channels for accelerated processing of data for anticipated signals from local / remote power generations, priority dispatching centers communication, priority technological signals from the control systems of large concentrated power generations can be used. This option is related to the control systems of the respective objects as well.

Some of the protection devices design related features shall be estimated for a "period in the future", covering the expected service time. The presence of elements with rotation components like cooling fans, HDDs, can shorten the service life and/or impede the necessity of servicing. Typically the "relay protection" system do not have such elements, but indirectly they can depend on the operation of cabinet ventilation/cooling, power supply units, common SCADA servers, etc. The life span of built-in batteries for real-time clock and other back-up functions can also impose limitations on long-term operation of the electronic devices.

Some of the listed functionalities for protection devices practically are not applicable for electromechanical and solid-state devices. Eventual comparison between electromechanical and microprocessor devices is applicable only for specific cases.

4 Conclusions

The proposed approach for estimation of digital protection devices applicability on basis of multiple characterizing parameters can be used for comparison of alternative options of protection schemes.

The proposed set of criteria for IED applicability estimation, if intended for practical application, shall be "tailored" to the particular grid /grid operator/ utility operator specifics. Some advanced grid protection techniques like application of Phasor Measurement Units (PMU) are still not adopted in many regions, but the swift technology development indicates positive tendency [13]. Such functionality can also be included in the "extended version" of the example given in Table 1. The balanced and perspective designed protection systems will reduce a lot of problems like black-outs, power interruptions, power quality indices, etc. It shall be counted, that applying more sophisticated equipment imposes the necessity of increased installation, maintenance and equipment physical protection practices [14].

On the basis of such approach a procedure can be developed for both technical and economical evaluation of equipment applicability [15]. The proposed frame of criteria (the included in Table 1 characteristics) is based on some of the basic characteristics of the IEDs/protection devices. Extending the scope of criteria may lead to sophisticated ranking approach for detailed review of the characteristics of high technological products. In the scope of criteria the respective relevant IEC/EN/IEEE standards can be included [16,17].

Such approach may support the selection of optimal components for protection system and in the same time to give a sound justification of the selection [18]. If protection devices are selected on the basis of compromise-their limited functionality, insufficient options for upgrade, software limitations, etc., may impose shortages for the operation, maintenance and enhancing the system functionality in the future.

References:

[1] Remus Teodorescu, M. Liserre, P. Rodríguez, Grid Converters for Photovoltaic and Wind Power Systems, 2011 John Wiley & Sons, Ltd. ISBN: 978-0-470-05751-3.

[2] Barkans J., Zalostiba D. Protection against Blackouts and Self-Restoration of Power Systems. RTU Publishing House, UDK 621.311.16, Riga, 2009.

[3] Learning from the Blackouts Transmission System Security in Competitive Electricity Markets, International Energy Agency (IEA) ISBN 92 64 10961, 7 December 2005.

[4] S. Nick, H. Wetzel, C. Growitsch, The Costs of Power Interruptions in Germany – an Assessment in the Light of the Energiewende, Institute of Energy Economics at the University of Cologne (EWI), Working Paper, No 13/07, April 2013.

[5] Interim Report: Causes of the August 14th Blackout in the United States and Canada. U.S.-Canada Power System Outage Task Force. November 2003.

[6] The insurance implications of a cyber-attack on the US power grid. Lloyd's Emerging Risk Report – 2015. OSCE. Organization for Security and Co-operation in Europe.

[7] TEIAS, ENTSO-E. Report on Blackout in Turkey on 31st March 2015. Final Version 1.0 Project Group Turkey, 21 September 2015.

[8] D. Novosel (chair) Performance of Relaying during Wide-Area Stressed Conditions, IEEE PSRC, WG C12, May 14, 2008.

[9] L. Shuran, D. Hui, G. Su. Analyses and Discussions of the Blackout in Indian Power Grid. Energy Science and Technology Vol. 6, No. 1, 2013, pp. 61-66.

[10] The Day the Sun Brought Darkness. NASA, https://www.nasa.gov/topics/earth/features/sun_darkness.html.

[11] J.W.Wang, L.L.Ronga, "Cascade-based attack vulnerability on the US power grid," Elsevier, Safety science, vol. 47, no. 10, pp. 1332–1336, 2009.
[12] Avi Schnurr. Vulnerability of national power grids to electromagnetic threats: domestic and international perspectives. Energy law journal. Volume 34, No. 1 2013.
[13] Mladen Kezunovic, S. Meliopoulos, V. Venkatasubramanian, V. Vittal. Application of Time-Synchronized Measurements in Power System Transmission Networks, 2014 Springer, ISBN 978-3-319-06217-4.
[14] C. J. Schrijver, R. Dobbins, W. Murtagh, and S. M. Petrinec. Assessing the impact of space weather on the electric power grid based on insurance claims for industrial electrical equipment. AGU Publications. Research article 10.1002/2014SW001066, published online 8 JUL 2014.
[15] Mediha Mehmed-Hamza. Research of the sensitivity of the selective instantaneous overcurrent protection relays for terminals 20 kV. Journal of the Technical University Sofia, branch Plovdiv, "Fundamental Sciences and Applications", Vol. 14, 2009, International Conference Engineering, Technologies and Systems, pp. 251-256 /in Bulgarian/.
[16] Organization for Security and Co-operation in Europe. Protecting Electricity Networks from Natural Hazards. 2016.
[17] ENTSO-E Subgroup "System Protection and Dynamics". Best Protection Practices For HV and EHV Transmission Systems of ENTSO-E CE Area Electrical Grids. Final version. 18 April 2012.
[18] ENTSO-E Subgroup "System protection and dynamics" Special protection schemes, March 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 205-212 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Design of an Automatic Forward and Back Collision Avoidance System for Automobiles

Tasneem Sanjana*, Ferdus Wahid, Mehrab Masayeed Habib, Ahmed Amin Rumel

Faculty of Electrical and Electronic Engineering, American International University-Bangladesh (AIUB), 1213, Bangladesh

ARTICLE INFO

Article history:
Received: 14 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
Forward collision avoidance system
Rear collision avoidance system
ATMEGA32
Sharp Distance Sensor
Motor driver
Actuator

ABSTRACT

This paper is the extended reflection of work originally presented in conference of Electrical, Computer and Communication Engineering (ECCE)-CUET 2017, entitled "Automated Anti Collision System for Automobiles". Automated collision avoidance system is a trending technology of science in automobile engineering. The aim of this paper is to design a system which will prevent collision from the front as well as the back for automobiles. This paper gives an overview of secure and smooth journey of car (vehicles) as well as the certainty of human life. This system is controlled by microcontroller ATMEGA32. Two Sharp distance sensors are used to detect object within the danger range where one is for front detection and other is for back detection. A crystal oscillator is used to produce the oscillation and generates the clock pulse of the microcontroller. An LCD and a GLCD are used to give information about the safe distance for front and rear respectively, and a buzzer is used as alarm. An actuator is used as automatic brake and inside the actuator there is a motor driver that runs the actuator. For coding "microC PRO for PIC" is used and "Proteus Design Suite Version 8 Software" is used for simulation.

1. Introduction

All of the greatest achievement of the science, automobile is the most probably the one, which significantly changed human life. A collision avoidance system is a current emerging technology in the field of automobile. It is also known as pre-crash system, forward collision warning system or collision mitigating or car anti-collision system. An automated car anti-collision system is an automobile safety system design to reduce the accidents as road traffic accidents are the largest cause of injury-related deaths worldwide .When driver himself is not concentrating on driving or any other parameters, that time it may cause damage to vehicle (car) as well as a life, that time automated ant-car collision device can play an effective role regarding the danger ahead. An automated anti-collision system device is placed within a car to warn its driver of any dangers that may lie ahead on the road. Once the detection is done this system either provide a warning to the driver when there is an imminent collision or take action autonomously without any driver command (by braking or steering both). This system runs by microcontroller that detects obstacle with speed sensor and stops the vehicle by giving instruction to the actuator. [1].

*Tasneem Sanjana, American International University, Bangladesh
Email: tsanjana909@gmail.com

2. Literature Review

In 2003, Honda introduced the first pre-crash system with autonomous braking. In 2003 Toyota added an "automatic partial pre-crash braking system" to the Celsior36. In 2006 Audi introduced "Braking guard" radar-assisted forward collision warning on AudiQ7. But these could not put a satisfactory performance. These all were not fully automated. In 2009, the U.S. National Highway Traffic Safety Administration (NHTSA) had begun studying to make frontal collision warning systems and lane departure warning systems. In 2012 a research by the Insurance Institute for Highway Safety examined how particular features of crash-avoidance systems affected the number of claims under various forms of insurance coverage. This indicates that two crash-avoidance features provide the biggest benefits; these are (a) autonomous braking that would brake on its own when the driver would not alert to avoid a forward collision and (b) adaptive headlights that would shift the headlights in the direction to the driver steers. At the circa 2012 stage of development, it has found from this research that lane departure systems is not helpful and perhaps harmful. Periodically anti-car collision features are rapidly making their way into the new vehicle fleet. Many car companies like BMW, Audi, Mercedes- Benz , Ford, Toyota, Volvo developed vehicle collision avoidance system from 2003-2009, but not necessarily they were all time effective.[1] Few years back,

some research were carried out on car anti-collision system device using ad hoc wireless network, V2V communication, GPS and radar implementation but all this effort were informatory in type which would give signal to the driver or produced some buzzes or sounds only but finally action would be taken by the driver own self in which there were quite chances of the collision. [2] Since the emergence of mechanical vehicles, road safety has been a major concern, the automated anti-car collision system has gained increasing attention in the last three years. As a result, various automated anti car collision system have originated for assuring safety than the existing system. Research of the Queensland department of Transport and Main Roads has revealed that installation of anti-collision technology could reduce 20 to 40 percent in terms of number and severity of fatal crashes and 30-50 percent of all injuries. Thus it is the main aim to ensure the safety of human life during the collision, the U.S National High Way Traffic Safety Administration has decided to make the pre-crash system mandatory for vehicles. In similar way all new vehicles in Europe will have advanced emergency automated braking system installed by November,2015. A group of automakers and suppliers companies including General Motor(GM), Ford Motor, Nissan Motor are doing research of this automated vehicles collision avoidance system at the University of Michigan. General Motor company has a plan to install this automated vehicles avoidance system in their cars by next two years. Toyota has already given an official announcement that they are going to give this automated vehicles avoidance system in their Toyota and Lexus models by 2017. On the other hand, BMW has already introduced satisfactory and effective anti-car collision system on BMW 7series in 2012. Volvo company has launched this system in Volvo FH by February 2013. Also car company Renault introduced automated active braking system on Renault EspaceV, in October 2014 as well as Volkswagen Passat introduced too on Volkswagen Passat/B8 in 2014. [2] [3] Today, automated anti car collision system generally has camera sensor, distance sensor, actuator, buzzer to detect and warn the drivers of any danger lie ahead on the road, it could be a car, human, pedestrian or any stationary object just as tree or pole etc.

3. System Overview

The block diagram of proposed system is given in Fig.1.Advantage of proposed block diagram over existing technologies is, this system will provide forward collision protection as well as helps to avoid rear collision by displaying messages to the vehicles coming from the back regarding to slow down their vehicle. Here microcontroller does the prime work and it controls the entire task. Two distance sensors are connected to the microcontroller. A LCD, A GLCD and an alarming device are also connected to the microcontroller. A Motor Driver is connected with micro controller to operate an actuator, as actuator needs high current to turn on and microcontroller can't provide this amount of current so motor driver will provide high current to actuator to turn on when it's needed. Distance sensors will be placed in front and back side of a car. They will sense the distance of vehicle in front and rear of the car and send this information to microcontroller. Microcontroller will compare this distance with some preset distances. If the measured distance from the front distance sensor crosses the preset safe distance value, microcontroller will send command to LCD to display close distance and also to write value of that exact distance, At the same time microcontroller will send a pulsating signal to buzzer and buzzer will be buzzed according to that signal. If the measured distance by the front distance sensor crosses the preset value of close distance in microcontroller, microcontroller will tell LCD to show critical distance and the value of that corresponding distance measured by the sensor.

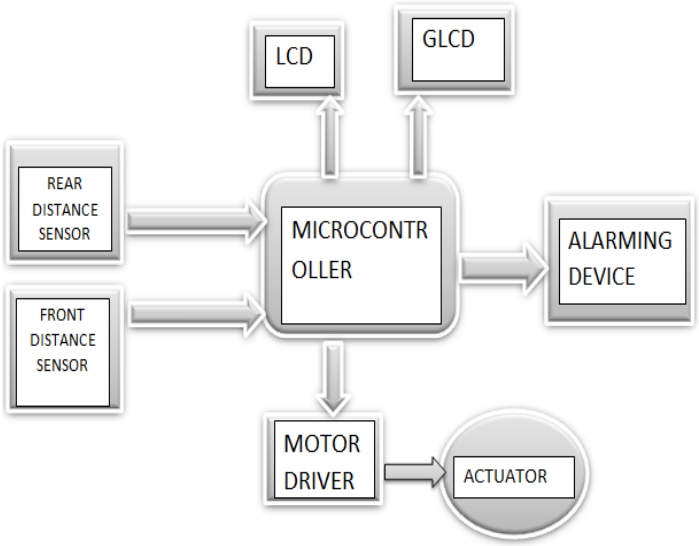


FIG.1. Block diagram of forward and back collision avoidance system for automobiles

At the same time a continuous signal will be send to buzzer and it will buzz continuously to notify the driver that vehicle is exceeding close distance and it's forwarding towards critical distance. Whenever vehicle crosses critical distance and entered into the distance at which collision is must if the car isn't slow down, that moment microcontroller will send signal to motor driver to turn on the actuator and it will stop the car automatically, at the same time LCD will "show actuator on" with distance information. Here, actuator itself works as a brake. For rear distance sensor it will sense the distance of the incoming vehicle and it will sent information to micro controller, if the distance exceed the safe distance value microcontroller will instruct GLCD to display slow down along with distance information, and if the incoming vehicle crosses critical distance microcontroller will instruct GLCD to display "Emergency! Stop" along with distance information to the driver of the incoming vehicle from the back. GLCD could be placed at the rear glass or just above the registration plate of the car containing this proposed safety technology.

4. Preparation to Proteus model

In this section components that were used for simulation and connection diagram has been explained.

4.1. ATMEGA32

ATMEGA32 is the high-performance, low-power Microchip 8-bit AVR RISC-based microcontroller combines 32KB ISP flash memory with read-while-write capabilities, 1KB EEPROM, 2KB SRAM, 54/69 general purpose I/O lines, 32 general purpose working registers, a JTAG interface for boundary-scan and on-chip debugging/programming, three flexible timer/counters with

compare modes, internal and external interrupts, serial programmable USART, a universal serial interface (USI) with start condition detector, an 8-channel 10-bit A/D converter, programmable watchdog timer with internal oscillator, SPI serial port, and five software selectable power saving modes. The device operates between 1.8-5.5 volts. By executing powerful instructions in a single clock cycle, the device achieves output through approaching 1 MIPS per MHz, balancing power consumption and processing speed. [4]

4.2. Sharp Distance Sensor: Model (GP2Y0A21):

A sensor is an electrical device which is used to measure physical properties and gives corresponding electrical output which is a transmitting impulse as for measurement or control for operation. In this model for measuring the distance of a vehicle Sharp Distance Sensor (GP2Y0A21) is used. Sharp Distance sensor follows triangulation measuring method for measuring distance of an object in front of it, It transmits an IR (Infrared Ray), if any obstacle is present in the path of that IR ray it will bounces back to the sensor and that's how distance is measured. Output of this sensor is analog and this output varies a range from 3.1V at 10cm to 0.4V at 80cm.

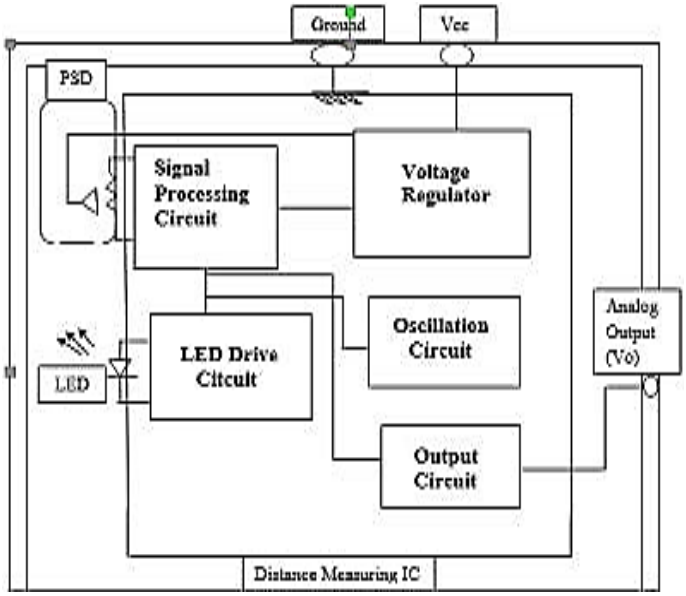


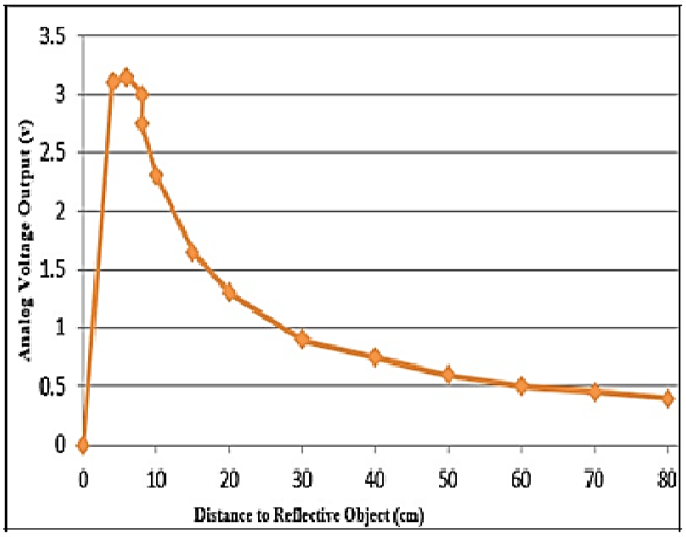
FIG.2.Internal Block Diagram of Sharp Distance Sensor [1]

FIG.2. indicates that this sensor contains IR LED armed with lens which discharges narrow light beam, after reflecting from the object, the beam will be directed through the second lens on a position-sensible photo detector (PSD) and the conductivity of PSD depends on the point where the beam falls, finally conductivity converts to voltage and voltage digitalizes by Analog to digital converter (ADC) [5]. Sharp Distance sensor utilize following equations to convert Distance into digitalized voltage with the help of ADC, and the relation between distance and the Voltage is

$$1 / (d + k) = a * ADC + b \qquad (1)$$

Where, distance is in centimeters.
k is corrective constant (fund using tial-and-error method)
ADC is digitalized value of voltage.
A is linear member (value is determined by the trend line equation)
b is free member (value is determined by the trend line equation).
So, distance d can be expressed from the formula [6]:

$$d = (1 / (a * ADC + B)) – k \qquad (2)$$



If a graph is plotted using equation (2) it will look like FIG.3. After analyzing this graph it can be concluded that the output voltage of the Sharp distance sensor is decreasing with the increased distance, which indicates an inversely proportional relation between distance and output voltage of sharp distance sensor.

4.3. Linear Actuator:

Linear actuator is a device which converts circular motion produced by a conventional electric motor into linear motion. Linear actuators are used in machine tools and industrial machinery, in computer peripherals such as disk drives and printers, in valves and dampers, and in many other places where linear motion is required. Hydraulic or pneumatic cylinders inherently produce linear motion. Many other mechanisms are used to generate linear motion from a rotating motor. In the majority of linear actuator designs, the elementary principle of operation is an inclined plane. The threads of a lead screw act as a continuous ramp that consents a small rotational force to be used over a long distance to complete the movement of a large load over a short distance. The system converts rotary motion (in the form of an electric motor) to linear motion. When an electrical source is applied to the motor, the thread shaft is driven and the nut rides up and down the shaft in the corresponding direction. This action delivers an extension and retraction capability for tasks requiring a linear displacement. Note that in this example the nut and red tube do not rotate, merely "ride" the threads on the spinning shaft up and down. [7]

4.4. Crystal Oscillator (16 MHz):

A crystal oscillator is an electronic oscillator circuit which uses in the mechanical resonance of a vibrating crystal of piezoelectric material to create an electrical signal with a very precise frequency and this frequency is usually used to keep track of time (as in quartz wristwatches), to provide a stable clock signal for digital integrated circuits, and to alleviate frequencies for radio transmitters and receivers, The most common type of piezoelectric resonator used is the quartz crystal, so oscillator circuits incorporating them became known as crystal oscillators but other piezoelectric materials including polycrystalline ceramics are used in similar circuits. [8]. A crystal is a solid in which the constituent atoms, molecules, or ions are packed in a repeatedly ordered, repeating pattern extending in all three spatial dimensions. Almost any object made of an elastic material could be used like a crystal, with appropriate transducers, since all objects have natural resonant frequencies of vibration. For example, steel is very elastic and has a high speed of sound. It was often used in mechanical filters before quartz. The resonant frequency depends on size, shape, elasticity, and the speed of sound in the material. High-frequency crystals are typically cut in the shape of a simple, rectangular plate. Low-frequency crystals, such as those used in digital watches, are typically cut in the shape of a tuning fork. For applications not needing very precise timing, a low-cost ceramic resonator is often used in place of a quartz crystal. When a crystal of quartz is suitably cut and mounted, it can be made to distort in an electric field by applying a voltage to an electrode near or on the crystal. This property is accepted as electrostriction or inverse piezoelectricity. When the field is removed, the quartz will generate an electric field as it returns to its previous shape, and this can generate a voltage. The result is that a quartz crystal behaves like a circuit composed of an inductor, capacitor and resistor, with a precise resonant frequency. Quartz has the further advantage that its elastic constants and its size change in such a way that the frequency dependence on temperature can be very low. The specific characteristics will depend on the mode of vibration and the angle at which the quartz is cut (relative to its crystallographic axes. Therefore, the resonant frequency of the plate, which depends on its size, will not change much, either. This means that a quartz clock, filter or oscillator will remain accurate. For critical applications the quartz oscillator is mounted in a temperature-controlled container, called a crystal oven, and can also be mounted on shock absorbers to prevent perturbation by external mechanical vibrations and oscillation.[8] In this design crystal oscillator is for to activate micro controller internal clock operating circuit.

4.5. Motor Driver (L293D):

L293D is a dual H-bridge motor driver integrated circuit (IC). Motor drivers act as current amplifiers since they take a low-current control signal and provide a higher-current signal. This higher current signal is used to drive the motors.L293D contains two inbuilt H-bridge driver circuits. In its common mode of operation, two DC motors can be driven simultaneously, both in forward and reverse direction. The motor operations of two motors can be controlled by input logic at pins 2 and 7 and 10 and 15. Input logic 00 or 11 will stop the corresponding motor. Logic 01 and 10 will rotate it in clockwise and anticlockwise. Enable pins 1 and 9 (corresponding to the two motors) must be high for motors to start operating. When an enable input is high, the associated driver gets enabled. As a result, the outputs become active and work in phase with their inputs. Similarly, when the enable input is low, that driver is disabled, and their outputs are off and in the high-impedance state. [9]

4.6. LCD Display (16x2):

LCD (Liquid Crystal Display) screen is an electronic display module and find a wide range of applications. A 16x2 LCD display is very elementary module and is very usually used in various devices and circuits. These modules are preferred over seven segments and other multi segment LEDs. The reasons is LCDs are economical, easily programmable, have no limitation of displaying special and even custom characters (unlike in seven segments), animations and so on. A 16x2 LCD means it can display 16 characters per line and there are 2 such lines. In this LCD each character is displayed in 5x7 pixel matrix. This LCD has two registers, namely, Command and Data. The command register stores the command instructions given to the LCD. A command is an instruction given to LCD to do a predefined task like initializing it, dissipating its screen, setting the cursor position, controlling display etc. The data register stores the data to be displayed on the LCD. The data is the ASCII value of the character to be displayed on the LCD. [10]

4.7. GLCD (AMPIRE128x64):

GLCD is a graphic liquid crystal display. AMPIRE (128*X*64) graphic LCD contains 128 coulombs and 64 rows. It can display data in (128x64) matrix. Graphical LCDs are different from the ordinary alphanumeric LCDS, like (16x1), (16x2), (16x4), (20x1), (20x2) etc. They (ordinary) can print only characters or custom made characters. They have a fixed size for displaying a character normally (5x7) or (5x8) matrix. Where as in graphical LCD we have 128x64=8192 dots each dot can be lit up as per coding or can make pixels with 8 dots that is 8192/8=1024 pixels, graphical LCD is controlled by two KS0108 controllers. A single KS0108 controller is capable of controlling 40 characters so for controlling a graphical LCD we need two KS0108 controllers and the 128x64 LCD is divided into two equal halves with each half being controlled by a separate KS0108 controller. [11] Special types of shapes or picture can be drawn in GLCD by exciting it DOTS.

4.8. MOSFET:

MOSFET is a three terminal semi-conductor device it consists of gate drain and source. It works in three region such as cutoff active and saturation. To use MOSFET as a switch it need to operate in active region and for amplification any input signal it will work in saturation region. Whenever a signal will be given in its gate it will activate. In this model MOSFET is used as a switch for alarming circuit (buzzer circuit). It's used in a way that voltage across its drain to source is close to zero to make it work as a switch.

4.9. Proteus Model:

FIG.4 is the proteus model of an automatic forward and backward collision avoidance system. This microcontroller works within 20nsec. Its operating voltage range is: 2.0V to 5.5V. Two

sharp distance sensors are used to measure the distance of the obstacle for front and rear respectively, crystal oscillator is also connected to the microcontroller, it produces the oscillation and generates the clock pulse of the microcontroller. There is a buzzer used as alarm and it is controlled by a MOSFET, microcontroller turns on the MOSFET which work as a switch for buzzer. A motor driver L293D is used for controlling the actuator. Actuator will be connected to the car braking system.

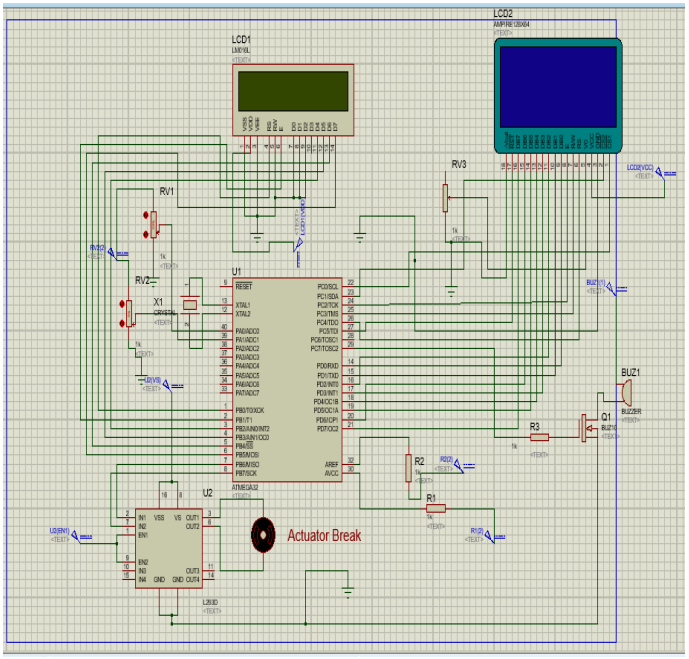


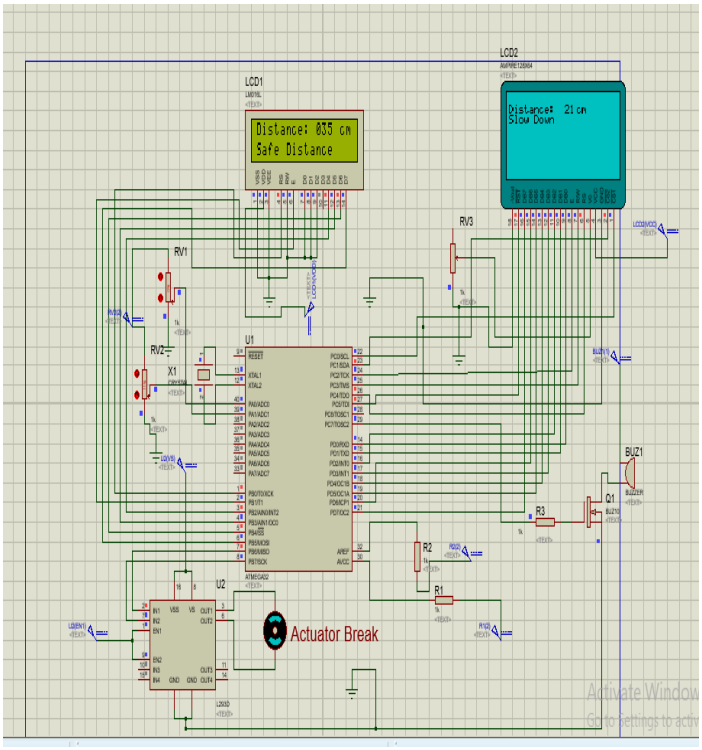
FIG.4. Proteus model of the system

From FIG.4 it is found that microcontroller pin7 (PB6/MISO) and pin8 (PB7/SCK) is connected to motor driver pin2 (IN1) and pin7 (IN2). From this connection of the configuration, motor driver gets command from microcontroller. 1k ohm resistors are connected to microcontroller pin30 and 32 with constant 5V DC for operating of microcontroller and ADC comparison respectively. ADC comparison is used in microcontroller because it will help microcontroller to process the measured distance by distance sensor in digital format. Microcontroller pin30 and 40 are connected to sharp distance sensors but there is no distance sensor device in Proteus so potentiometers are used instead of it. The highest readings of these potentiometers are 5k ohm and operating voltage +2.5V. Two port of actuator is connected to motor driver pin3 (OUT1) and pin6 (OUT2) A DC gear motor is used as an actuator in simulation for analyzing the motor driver behavior to actuator due to absence of an actuator in proteus model. Two ports of crystal oscillator are connected to microcontroller pin13 and pin12 is used for generating clock pulse of microcontroller. Here is a LCD display which pin11 (D4), pin12 (D5), pin13 (D6), pin14 (D7) are connected to microcontroller pin3 (PB2), pin4 (PB3/PGM), pin5 (PB4), pin6 (PBS) respectively, so that, LCD display can get signal from microcontroller and gives information about the obstacle.

Microcontroller pin29 (PC7) is connected to a MOSFET by a 1kohm resistor R2. Microcontroller itself cannot turn the buzzer on for this reason a MOSFET has been used. Microcontroller gives instruction to the MOSFET to turns the buzzer on and it operates at +12V. GLCD pin9-16 (DB0-DB7) is connected with microcontroller pin14-21 (PD0-PD7) which will give information regarding rear obstacle distance. GLCD control pins, 1(CS1), 2(CS2), 6(RS), 7(RW), 8(E), 17(RST) is connected to microcontroller pin22 (PC0),pin23 (PC1), pin26 (PC4),pin25 (PC3), pin24 (PC2),pin27 (PC5) respectively, so that GLCD can get signal from microcontroller and gives information about the distance which should be kept to avoid collision, or to stop to avoid collision towards the car approaching from the back. A program is written in MIKROC software, which is the instruction for microcontroller action and HEX file of that program is used for simulation.

5. Simulation and Result

From FIG.5.we can conclude that if there is a vehicle in front of a car containing this collision avoidance system at a distance of 35 cm microcontroller will command LCD to Display “safe distance “to the driver of the car, and at the same time if there is a vehicle approaching from rear and if it’s distance is 21 cm micro controller will instruct GLCD to display the distance information along with the message “Slow Down’’ to the incoming car from the back.



From FIG.6. we can conclude that if there is a vehicle in front of a car containing this collision avoidance system at a distance of less than 26 cm microcontroller will command LCD to Display “Close Distance” to the driver of the car and microcontroller will send activation signal towards buzzer circuit, and at the same time if there is a vehicle approaching from rear and if it’s distance is less than 15 cm micro controller will instruct GLCD to display the distance information along with the message “Emergency! Stop’’ to the incoming car from the back.

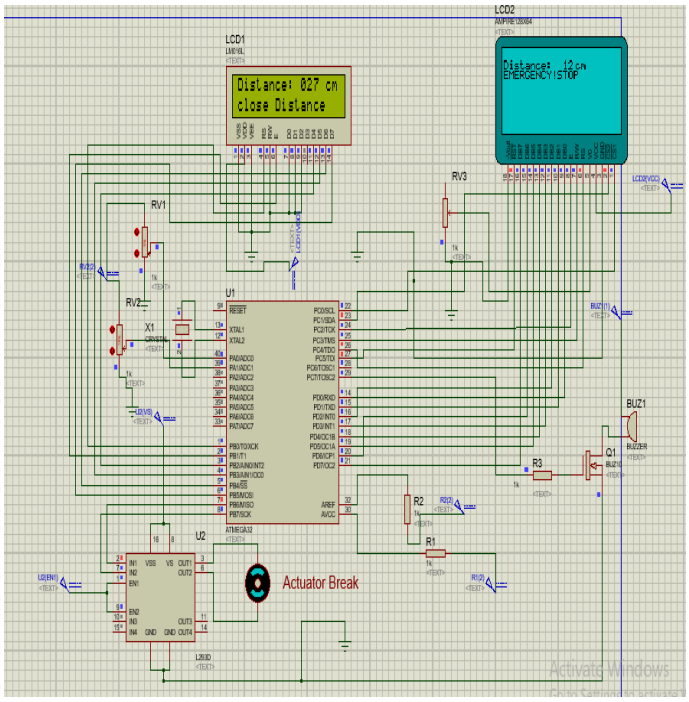


FIG.6. Result for close distance for front and emergency stop distance for rear

From FIG.7. shows that if there is a vehicle in front of a car containing this collision avoidance system at a distance of 18 cm microcontroller will command LCD to Display "Critical Distance" to the driver of the car and microcontroller will send activation signal towards buzzer circuit, and at the same time if there is a vehicle approaching from rear and if it's distance is27 cm micro controller will instruct GLCD to display the distance information along with the message "Maintain This Distance'' to the incoming car from the back.

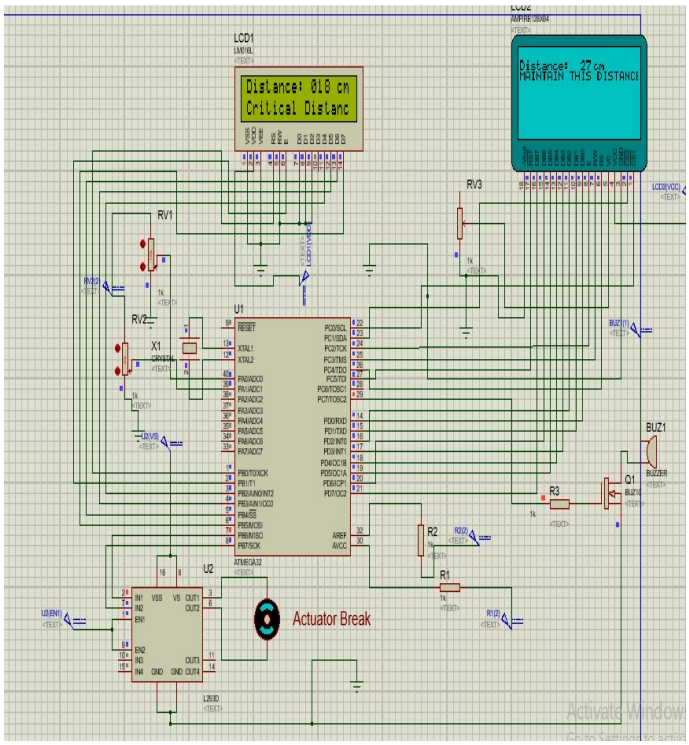


FIG.7. Result for critical distance for front and safe distance for rear

From FIG.8. we can conclude that if there is a vehicle in front of a car containing this collision avoidance system at a distance of 13 cm microcontroller will command LCD to Display "Actuator On" to the driver of the car and microcontroller will instruct motor driver to activate actuator, motor driver will send high current towards actuator which will cause actuator to turn on and this actuator will automatically press the brake of the car to stop. If there is a vehicle approaching from rear and if it's distance is less than 19 cm micro controller will instruct GLCD to display the distance information along with the message of "TENT TO COLLISION'' to the incoming car from the back.

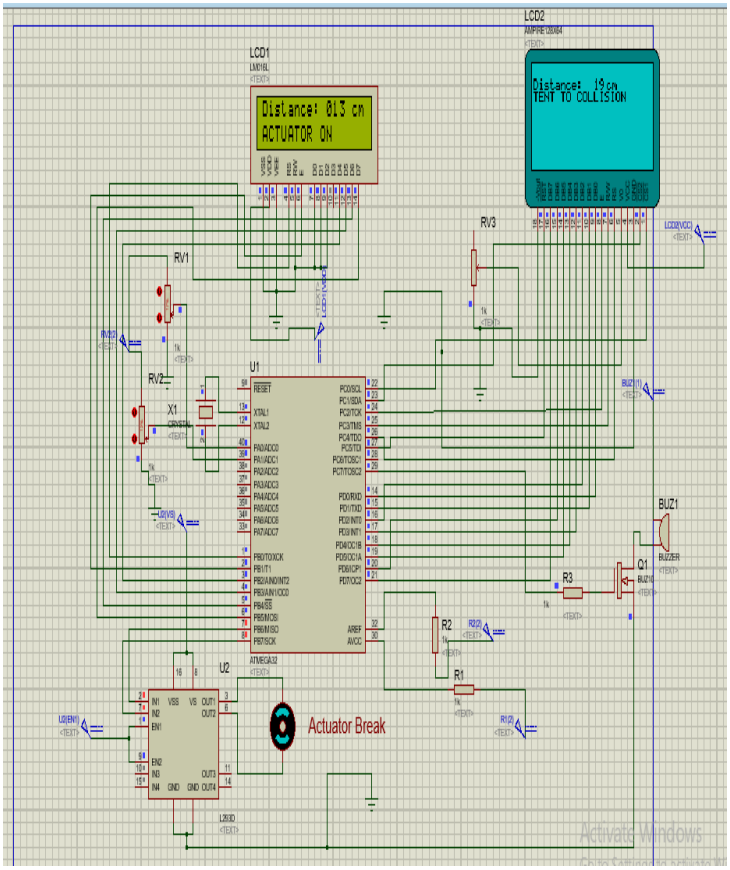


FIG.8. Result for Actuator On distance for front and tent to collision distance for rear

6. Result analysis:

During simulation distance was taken in small scale as in cm. For forward collision prevention system instruction for microcontroller was set in a way if measured distance by distance sensor is above 30 cm it will consider this distance to safe distance and it will instruct LCD to display to show the message safe distance along with that distance value. If the value of distance varies from 20 cm to 29 cm microcontroller will instruct LCD to show Close distance along with the value of the distance measured by the sharp distance sensor, further more microcontroller will send a pulsating signal to buzzer and buzzer will be turn on, this pulsating signal will cause change in the sound frequency of that buzzer. And if distance ranges from 15cm to 19 cm micro controller will instruct LCD to display the message "critical distance along with the measured distance "at the same time microcontroller will send a continuous signal to the buzzer and it

will buzzed, this buzzing frequency will be different from the previous one. And any distance less than 15 cm microcontroller will instruct motor driver to turn on the actuation system to stop the car. At the same time LCD will display "Actuator on along with distance information". From the simulated result, which were shown in the previous section; it is seen that when sharp distance sensors measure a distance of 35 cm LCD shows the message of safe distance which is larger than 30 cm. When measured distance by sensor was found 27 cm LCD shows the messages that this distance is close distance and at the same time a pulsating signal was sent to buzzer circuit. This measured distance was less than 29 cm. When distance was found about 18 cm LCD display message that this distance is critical distance, which was less than 20 cm. At the same time buzzer circuit will be turn on as micro controller sends a continuous signal to buzzer. From FIG.7 it can be seen that a red dot is appearing in front of the R3 resistor and which was connected with the microcontroller PC7 port and same red dot is appearing in that particular pin which indicates that microcontroller sending continuous signal to buzzer to buzz. Finally when distance was measured 13 cm which is less than 15 cm LCD displays the message that actuator on along with the measured value of distance in cm. From FIG.8 it is seen that there were two red dots in pin 2 and 7 of the motor driver which were connected to microcontroller pin 7 and pin 8 and same dots appear across those pins, which indicates microcontroller is sending high signal to motor driver to turn on the actuator or the motor of the actuator system. As motor driver needs high or low logic in its both pin to stop the motor of actuating signal and microcontroller is sending high signal to that motor driver is evidence that microcontroller is actually control the braking system of the vehicle and it will cause the car to stop.

For back collision avoidance system, the set of instruction for microcontroller was if measured distance is above 25cm it will be considered as a safe distance and GLCD will show "Maintain this distance" along with measured distance. Any distance less than 25 cm and greater 20 cm microcontroller will instruct GLCD to show "slowdown" along with measured distance value. Any distance ranges 19 cm to 15 cm microcontroller will instruct GLCD to display "tent to collision" along with the distance that was measured by sharp distance sensor. Finally any distance measured by the sharp distance sensor less than 15 cm microcontroller will instruct GLCD to display "Emergency! Stop". FIG.5 to FIG.8 in the simulation and result section it's seen that the measured distance by rear sharp distance sensor was 21cm, 12c m, 27 cm, 19 cm respectively, this distances follow the ranges that were mentioned above . And GLCD is displaying same sort of the message for that particular range of distances. From the above discussion it can be concluded that microcontroller is executing its instruction perfectly as it was given to it by a program which was written in MIKROC PRO for AVR, which results in that proposed design system is perfectly working during simulation. But when this system will be practically implemented for vehicle this distance range wouldn't be applicable, virtually all vehicles' or cars' road braking performance test indicate stopping distances for 60 mph are typically 36m to43m [12]. So when this system will be implemented in reality only the set of instructions will need to be changed which will be loaded in microcontroller.

7. Implemented forward collision system

The design of the research work in this paper is the extended design of the implementation of *"Automated Anti Collision System for Automobiles"* [1] which has been implemented successfully in practical by prototype prior, where a Sharp distance sensor, microcontroller PIC16f876a, motor driver, linear actuator alarming device and LCD display were used. From following figures we will see the complete view of previously implemented prototype device along with its distance ranges.

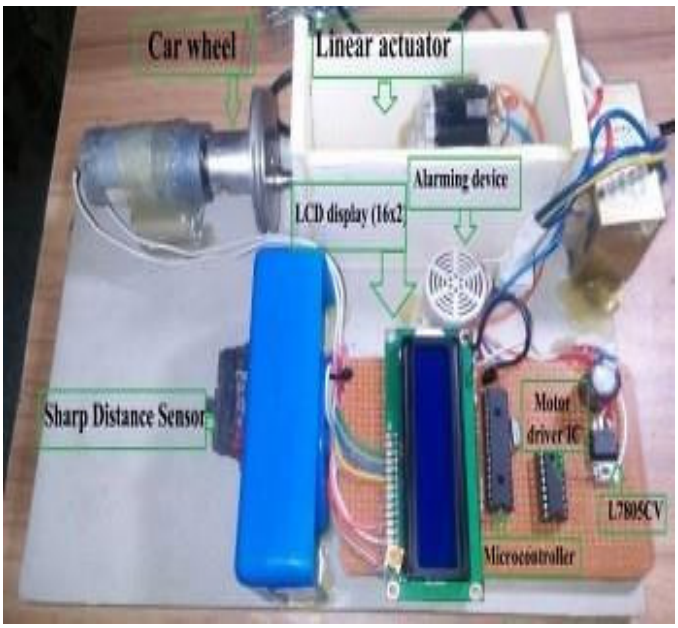


FIG.9: Implemented Prototype Device

From FIG.9.we can see the complete view of previously implemented prototype device.

FIG.10. Safe distance of the prototype device

From FIG.10.we can see the distance is 30 cm which was considered as safe distance and the actuator did not work. In FIG.11.and FIG.12.we can see distances as follow 24 cm and 09 cm whereas, 21 cm gave alarm to the actuator and in 09 cm, and actuator started working. In addition, the research of this paper has been done with the back collision avoidance system along with this mentioned forward collision avoidance system. Also more progress has been done about the forward obstacles detection distance ranges. Due to shortage of time this recent extended research work prototype was not possible to implement.

8. Application and Improvement plan of the system

Proposed system which is described in this paper will be implemented by prototype practically very soon. Near future, during manufacture it can be implanted in motor bikes taxi-cab, car, and truck even in aircraft system. In winter when roads might be rarely seen due to fogs this system will notify the person who will be in control of that vehicle about other vehicles or any other obstacles ahead of his vehicle. Which will help him drive safely his vehicle, and it will make him feel comfortable while driving and it will also extinguish his fear. This will reduce terrifying collision of the vehicles and not only the winter, in other season it will notify driver to maintain a safe distance by alarming sound or showing distance information in LCD. If driver begin to avoid the alarming sound and start to drive recklessly system will forcefully stop the car to avoid accident. Researchers of this paper has some future improvement plan for making this system more convenient and add other features to make this system more accurate. For improving this system sonar system will be implemented due to its long range detection, Emergency light or special kind of alarm will be used to notify the vehicle to maintain safe distance, graphical symbols will be added along with messages in GLCD. Wireless Fidelity (Wi-Fi) technology can be used for transferring information between successive vehicle about the distance and what speed should the kept for safety, furthermore, Dual Infra-Red Obstacle Detector (DROD40) can be used in this system to determine obstacle in every side such as left, right and ahead distinctly. In future, researchers of this paper have plan to design a system which will monitor driver health condition to judge his fitness for driving if he isn't fit, vehicle will not start. So, this will reduce almost 60% of accident.

Conclusion

Technologies, new inventions and transport system play an important role in human lives. Mankind is becoming habituated to these new inventions. Especially transport system without this mankind even think about their daily activities and even can step outside from their home. As every other invention this transport system has adverse effect too, and as usual mankind is responsible for this kind of adverse effect. People are always in rush to reach their destination as early as possible to save their time or to reach their destination on time without thinking about their lives value, what would happen if they met up with an accident which means colliding with other vehicle for their tendency to go somewhere within a flash. This results in injury or loses of their lives. The prime motive of the design is to reduce these accidents. In other words it can be concluded that this design is used to save human lives from the loss due to accident and also to ensure safe and comfortable journey.

Acknowledgment

First and foremost, we want to express our gratitude towards Almighty creator for his blessings without which it would not be possible to complete this work successfully afterwards want to acknowledge the constant support received from our family. It is an honor to express our heartiest gratitude to MS. Muhit, Assistant Professor, Faculty of Electrical and Electronic Engineering, AIUB, for his immense support and encouragement to improvise this designed system. And finally especial gratitude to honorable Mr. Chowdhury Akram Hossain, Assistant Professor, Faculty of Electrical and Electronic Engineering, AIUB. Without his motivation this work wouldn't be possible to advance so far. Finally, any errors are our own and should not tarnish the reputations of these esteemed persons.

References

[1] Tasneem Sanjana, Kazi Ahmed AsifFuad, Mehrab Masayeed Habib, Ahmed Amin Rumel "Automated Anti-Collision System for Automobiles" in International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar Bangladesh February 16-18, 2017 http://**DOI:** 10.1109/ECACE.2017.7913024

[2] http://www-nrd.nhtsa.dot.gov/pdf/esv/esv19/05-0322-O.pdf [Cited: 14 May, 2015]. Available:

[3] N., Mechanisms and Mechanical Devices Source book 4th Edition (2007), 25, McGraw- Hill [Cited: 17 May 2015].

[4] http://www.microchip.com/wwwproducts/en/ATmega32 [Cited: 13 May, 2015]. Available:

[5] http://www.sharpsma.com/webfm_send/1208 [Cited: 17 May, 2015]. Available:

[6] Roboticlab website abailable at, http:// home. roboticlab.Eu/en/examples/sensor/ir_distance(Cited : 17, May 2016).

[7] https://www.firgelliauto.com/blogs/news/18090539-linear-actuator-how-to[Cited: 17 May, 2015]. Available:

[8] http://www.electronics-tutorials.ws/oscillator/crystal.html[Cited: 17 May, 2015]. Available

[9] http://www.engineersgarage.com/sites/default/files/L293D.pdfCited: 7 May 2015]. Available:

[10] http://www.engineersgarage.com/sites/default/files/LCD%2016x2.pdf [Cited: 17 May , 2015]. Available:

[11] https://www.electronify.org/basic-component-introduction/graphics-lcd [Cited: 17 May, 2015]. Available:

[12] Divya P.&MurugesanA.,"Intelligent Car Braking System with collisinAvoidance and ABS",International Journal of Computer Applications(0975– 8887),National Conference on Information and Communication Technologies (NCICT , 2015).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 213-217 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A High Efficiency Ultra Thin (1.8 μm) CdS/CdTe p-i-n Solar Cell with CdTe and Si as BSF layer

Nahid A. Jahan, Md. Minhaz Ul Karim and M. Mofazzal Hossain*

Department of Electronics and Communications Engineering, East West University, Dhaka 1212, Bangladesh

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords :
CdTe/CdS
BSF layer
Cell efficiency
Intrinsic layer

ABSTRACT

CdTe-based photovoltaic (PV) cells provide the lowest EBPT (energy payback time) and emit less amount of GHG (green house gases) among different types of PV cells. Thus, it is very essential to enhance the efficiency of CdTe-based solar cells. A high efficiency CdTe/CdS p-i-n heterostructure solar cell is designed and the performance of the cell is investigated. All the simulations have been done using AMPS 1D (Analysis of Microelectronic and Photonic Structures) simulator. The cell consists of a transparent conducting oxide (TCO) layer (ZnO), a window layer (n-type CdS), an intrinsic layer (CdTe), an absorber layer (p-type CdTe) and a highly doped CdTe and Si as back surface field (BSF) layer. After the optimization of layer thicknesses and doping densities of different layers, we obtained a conversion efficiency of 26.01% for a total cell thickness of 1.8 μm. It is also found that the conversion efficiency can be increased by simply increasing the thickness of intrinsic layer. At 1.5 AM solar irradiance, the proposed cell structure attained a V_{oc} of 1.09 V, a J_{sc} of 26.78 mA/cm^2, and a fill factor of 89%, reaching an overall conversion efficiency of 26.01% with CdTe as BSF layer.

1. Introduction

This paper is an extension of our previous work presented in *ICAEESE 2016* [1]. The rapidly rising demand of world energy supply and the depletion of fossil fuel impose crucial challenges concerning today's energy requirement context. It is also worth mentioning that, the surfeit burning up of fossil fuels renders threatening global warming emissions (GWE) ascribing alarming rise of terrestrial temperature while imposing serious environmental pollutions by particulate materials [2]. These aforementioned issues necessitate the blooming of alternatively dependable, affordable-cost consistent energy sources. To meet these requirements various plausible renewable energy technologies have been thriven so far. Among these alternative technologies solar photovoltaic (PV) is thought to be as one of the most potential and reliable green energy concepts that assures the direct conversion of incident solar irradiation into electrical energy. The power extracted from PV architectures is popularly being adopted and deployed in the remote areas while ensuring the minimization of electrical power transmission losses. PV power can also be potentially applied in vehicular transportation that can ensure the diminishing of global warming emissions in near future. From purely economic point of view, the flourishing market of PV is generating a shortfall in the storage of highly cost silicon wafers which takes almost 50% of the total manufacturing cost. Therefore, it fosters an opportunity for thin film solar technology to create credentials in the market of PV modules because of its moderate cost. Among different types of thin film based modules, CdS/CdTe thin-film solar cells reveals significant suitability owing to its distinctive characteristics encompassing the following aspects. Firstly, the energy bandgap of CdTe is 1.45 eV, which is favorably well-matched to the spectrum of solar radiation. It also lies in the category of direct bandgap semiconductor guiding to fairly strong light absorption with reduced generation of heat. Besides, the doping of CdTe as p-type semiconductor film as well as the formation of p-n heterojunction with the inclusion of CdS is found to be much easier. Moreover, CdS exhibits wide energy-gap of 2.4 eV and can be fabricated as n-type adopting undemanding usual techniques of film deposition [3-5]. In the meanwhile, various less-costly and manageable atomic congregation techniques have been developed for the growth of CdS/CdTe solar cells [6-7]. Also it should be mentioned that, a thickness of around 4 microns absorber layer of CdTe is almost appropriate for capturing the entire solar spectrum due to acceptable range of absorption coefficient [8-9] of CdTe (~10^5 cm^{-1}) in the visible solar spectrum. All these properties

*M. Mofazzal Hossain, Aftabnagar, Dhaka-1212, Bangladesh, +8801796587888 & Email: dmmh@ewubd.edu

of CdTe attribute to quite less material cost of the cell if compared with cells based on Si-wafers. In addition, from various respective studies, it is revealed that the energy-payback-time (average) of thin film CdS/CdTe solar cell is around 1.4 years [10] that is found to be as the rock-bottom among all PV technologies. Moreover, CdTe solar cell generates the lowest amount of GWE for per kWh electrical energy compared to other PV technologies [10].

In CdS/CdTe thin film solar cell, irrespective of superstrate or substrate configurations, light enters into the cell through the TCO and CdS (n-type) window layers. From the earlier attempts, the experimental efficiency of single junction CdS/CdTe solar cell has been reported so far around 17% [11-13]. Approximately 18.6% conversion efficiency from a CdS/CdTe solar cell incorporating complex BSF and TCO layers has been reported by Amin et al [14].

In this work, we explore the possibilities and chances of enhancing the efficiency of CdS/CdTe heterojunction solar cell with the insertion of an intrinsic layer of CdTe. Additionally, a back surface field (BSF) layer of highly doped CdTe (p^+ layer) and Si (p^+ layer) are sandwiched between back contact and the absorber layer of p-CdTe in order to diminish the recombination loss and thereby, magnifying the conversion efficiency. In this proposed structure as depicted in Figure 1, a TCO layer of ZnO is inserted to enhance the level of V_{oc} by minimizing the series resistance of the cell.

2. Methodology

In this work the proposed cell is simulated in AMPS 1D simulator to optimize the layer thicknesses of TCO, window, absorber, intrinsic and BSF layers, and the doping density of BSF layer to achieve the maximum possible conversion efficiency under 1.5 AM solar irradiance. The properties of n-CdS, p-CdTe, a-Si and n-ZnO, those are given in Table 1, are taken from [13-15].

Table 1 Material Properties

Parameters	a-Si	CdTe	CdS	ZnO
ε_r	11.9	9.4	9.0	9.0
μ_n ($cm^2.V^{-1}.s^{-1}$)	20	500	350	100
μ_p ($cm^2.V^{-1}.s^{-1}$)	2	60	50	25
E_g (eV)	1.72	1.45	2.42	3.0
N_C (cm^{-3})	2.5×10^{20}	8×10^{17}	2.4×10^{18}	2.2×10^{18}
N_V (cm^{-3})	2.5×10^{20}	1.8×10^{19}	1.79×10^{19}	1.8×10^{19}
χ (eV)	3.8	4.28	4.5	4.35

3. Function of Intrinsic Layer

To comprehensively explain the function of intrinsic layer the simplified energy band diagram for thermal equilibrium condition is shown in Figure 2. It is worth mentioning that while depicting the energy band diagram, gradient of the quasi-Fermi level has not been shown for simplicity. Literally, when light falls and photo-voltage is generated it biases the p-n junction in forward bias mode and therefore the Fermi level should not be continuous throughout the junction of cell. However, it is evident from this diagram that, the inclusion of intrinsic layer increases the width of depletion layer where mainly the photons are absorbed and electron-hole pairs are generated. From Figure 2, it is clear that the generated electrons can move easily towards the TCO layer (left to n-CdS) due to the slope of electron energy and the effect of tunneling. Similarly photo-generated holes can move towards the back contact layer (right to BSF layer) due the slope of holes energy.

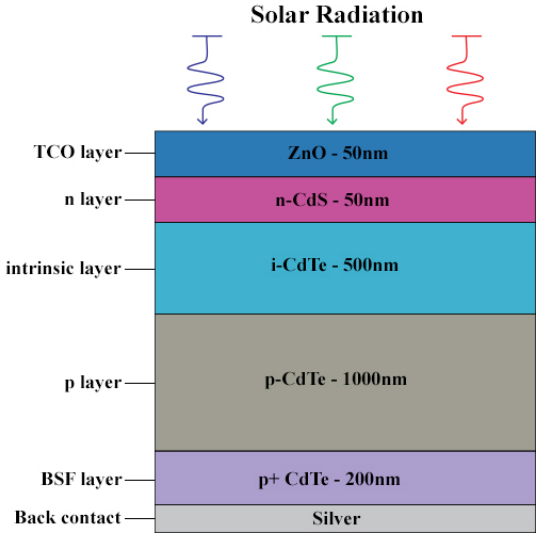


Figure 1The structure of the proposed p-i-n CdS/CdTe solar cell

4. Results and Discussion

In this work, using AMPS 1D simulator we have optimized the thickness of different layers of the cell in order to improve the efficiency. Throughout the simulation, the thicknesses of window layer (n-CdS) and absorber layer (p-CdTe) are kept constant at 50 nm and 1000 nm respectively (as shown Figure 1). We only explored the possibilities of enhancing the conversion efficiency of the cell by optimizing the thicknesses of BSF and intrinsic layers. The proposed structure is simulated in AMPS 1D environment and the I-V characteristics of the cell under 1 sun illumination is extracted from AMPS 1D simulator which is exemplarily shown in Figure 3. After simulation under different doping densities and thicknesses we obtained several figures and results from which we extracted the cell efficiency for desired doping density and thickness.

We explored the effects of doping density of p-CdTe BSF layer on the conversion efficiency for a BSF layer thickness of 50 nm and the results are shown in Figure 4. From Figure 4 it is found that beyond doping density of 2×10^{18} cm^{-3}, the conversion efficiency increases exponentially. The effects of the insertion of amorphous Si (a-Si) and CdTe as BSF layer on the efficiency are also investigated.

Figure 5 and Figure 6 depict the effects of BSF layer thickness (highly doped p-Si and p-CdTe) on the conversion efficiency without the inclusion of intrinsic layer. For this analysis, the doping density of a-Si and CdTe are considered as 6.0×10^{17} cm^{-3} and 1.0×10^{19} cm^{-3} respectively . At a BSF layer thickness of 200 nm, the maximum efficiencies for a-Si and CdTe are attained to be 23.0% and 24.46% respectively. It is observed that beyond 200 nm thickness of BSF layer, in case of a-Si the efficiency becomes

saturated whereas in case of CdTe the efficiency increases almost linearly. From Figure 5 and 6 it is also noticed that in the proposed CdS/CdTe solar cell, CdTe outperforms a-Si as a BSF layer.

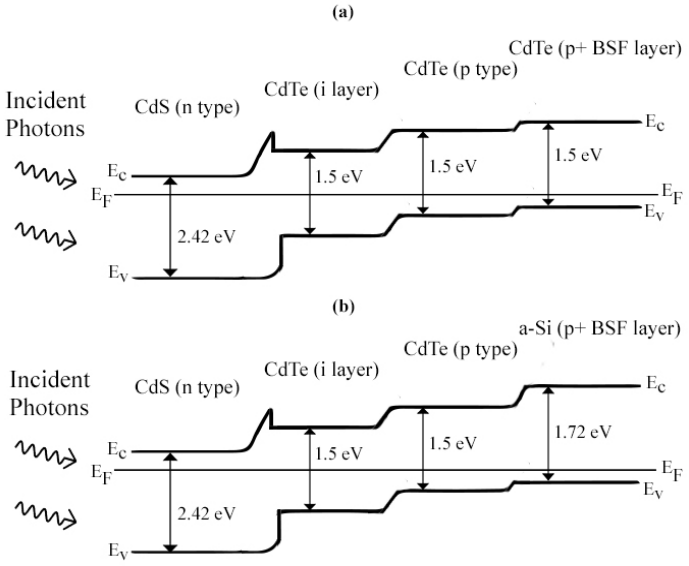


Figure 2 Simplified energy band diagram of proposed (a) p-i-n CdS/CdTe (BSF layer: CdTe) (b) p-i-n CdS/CdTe (BSF layer: a-Si) solar cells

Figure 7 and 8 show the effects of intrinsic layer (CdTe) thickness on the cell's conversion efficiency considering a-Si and CdTe as BSF layers separately. It is observed that in both cases the conversion efficiency increases linearly with the increase of intrinsic layer thickness. However, the maximum conversion efficiency at thickness 500 nm of intrinsic layer for CdTe and a-Si (as BSF layer) are found to be 26.01% and 24.51% respectively. The dependence of short circuit currents on the variation of intrinsic layer thickness for both CdTe and a-Si as BSF layers are also studied as depicted in Figure 9. From this comparison it is realized that CdTe as BSF layer provides larger short circuit current than that of a-Si.

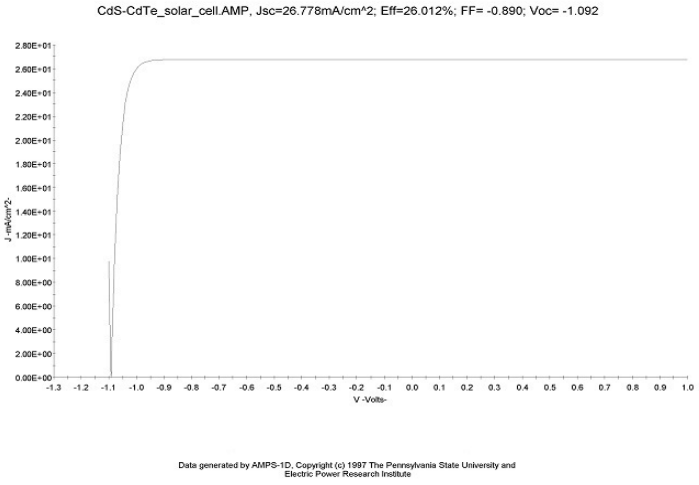


Figure 3 I-V characteristics of the proposed p-i-n CdS/CdTe solar cell.

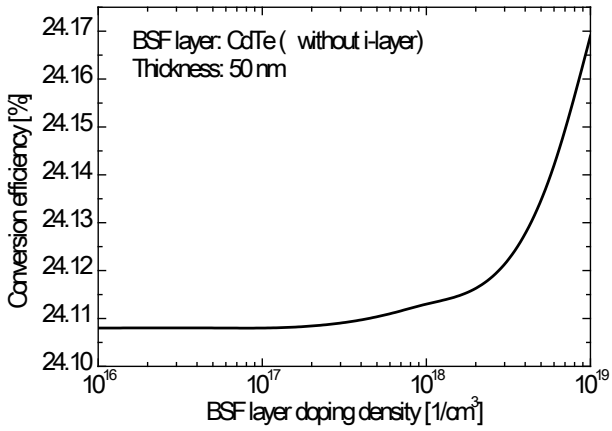


Figure 4 Effects of BSF layer doping density on the conversion efficiency

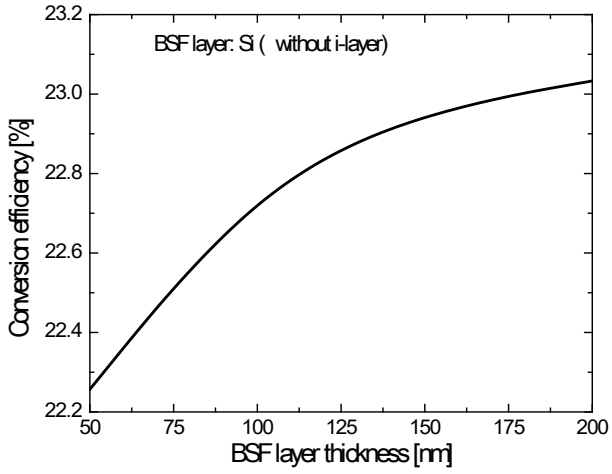
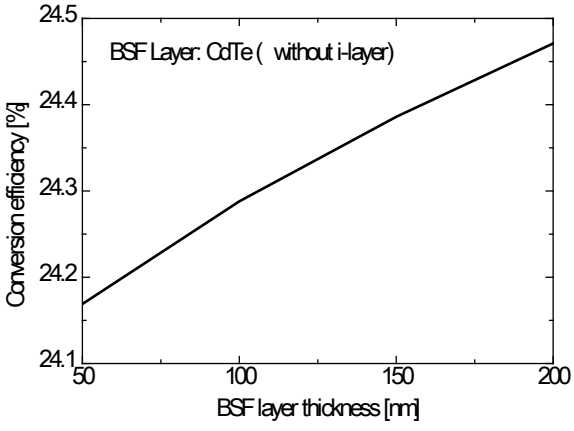


Figure 5 Effects of BSF (Silicon) layer thickness on the conversion efficiency



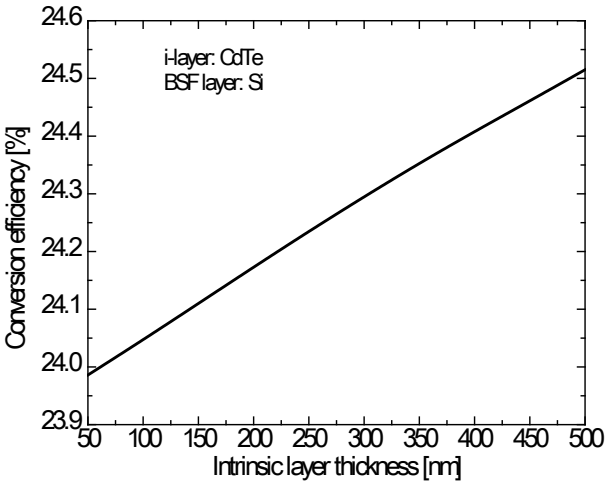


Figure 7 Effects of thickness of intrinsic layer on the conversion efficiency

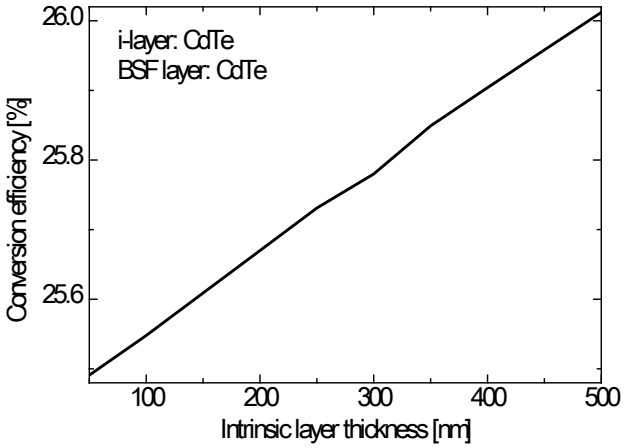


Figure 8 Effects of thickness of intrinsic layer on the conversion efficiency

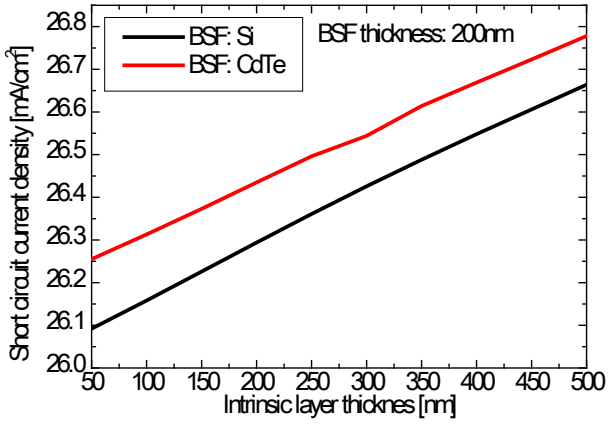


Figure 9 Comparison of current density for Si and CdTe as BSF layer

5. Conclusions

An ultrathin (1.8 μm) p-i-n CdS/CdTe PV cell is designed and simulated using 1D AMPS simulator. The thicknesses of intrinsic layer and BSF layer and the doping density of BSF layer are optimized. It is found that CdTe outperforms Si as a BSF layer. The maximum conversion efficiency of the proposed solar cell is found to be 26.01% taking CdTe as BSF layer. Simulation results revealed that a highly doped p^+ CdTe layer (BSF layer) has significant contribution in achieving such a high efficiency. At 1.5 AM solar irradiance, the proposed cell structure with CdTe as BSF layer achieved an open-circuit voltage of 1.09 V, a short-circuit current density of 26.78 mA/cm^2, and a fill factor of 89% and the corresponding overall conversion efficiency of 26.01%. The same structure with Si as BSF layer achieved an open-circuit voltage of 1.06 V, a short-circuit current density of 26.66 mA/cm^2, and a fill factor of 87%, and the corresponding overall conversion efficiency of 24.51%. Compared to our previous work [1] the cell thickness is reduced by 30% with a sacrifice of 0.73% cell efficiency.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors are thankful to East West University, Dhaka, Bangladesh for financial supports.

References

[1] M. Mofazzal Hossain, Md. Minhaz Ul Karim, S. Banik, Nahid A. Jahan and M. A. Matin, "Design of a high efficiency ultrathin CdTe/CdS p-i-n solar cell with optimized thickness and doping density of different layers," *ICAEESE*, Putrajaya, Malaysia 2016, 14-16 Nov.
DOI: 10.1109/ICAEES.2016.7888058

[2] A. Nishimura, Y. Hayashi, K. Tanaka, M. Hirota, S. Kato, M. Ito, K. Araki and E.J. Hu "Life cycle assessment and evaluation of energy payback time on high-concentration photovoltaic power generation system," Applied Energy **87**, pp. 2797-2807, 2010.
https://doi.org/10.1016/j.apenergy.2009.08.011

[3] G. C. Morris and S. K. Das, "Some fabrication procedures for electrodeposited CdTe solar cells," International Journal of Solar Energy, **12**(1-4), pp. 95-108, 1992.
http://dx.doi.org/10.1080/01425919208909753

[4] T. L Chu, "Thin film cadmium telluride solar cells by two chemical vapor deposition techniques," Solar Cells, **23**, pp. 31-48, 1988.
https://doi.org/10.1016/0379-6787(88)90005-1

[5] S. Ikegami, "CdS/CdTe solar cells by screen-printing-sintering technique: Fabrication, photovoltaic properties and applications," Solar Cells, **23**(1-2), pp. 89-105, 1988.
https://doi.org/10.1016/0379-6787(88)90009-9

[6] T. L Chu and S. S. Chu, "High efficiency thin film CdS/CdTe solar cells," International Journal of Solar Energy, **12**(1-4), pp. 121-132, 1992.
http://dx.doi.org/10.1080/01425919208909755

[7] V. M. Nikale, S. S. Shinde, C. H. Bhosale, and K. Y. Rajpure, "Physical properties of spray deposited CdTe thin films: PEC performance," Journal of Semiconductors, **32**(3), pp. 033001-7. 2011.
http://doi.org/10.1088/1674-4926/32/3/033001

[8] Marple, D. T. F., "Optical Absorption Edge in CdTe: Experimental," Phys. Rev. **150**(2), pp. 728-734, 1966.
https://doi.org/10.1103/PhysRev.150.728

[9] Kim Mitchell1, Alan L. Fahrenbruch and Richard H. Bube, "Photovoltaic determination of optical-absorption coefficient in CdTe," J. Appl. Phys. **48**, pp. 829-830, 1977.
https://doi.org/10.1063/1.323636

[10] Jinqing Peng, Lin Lu, and Hongxing Yang, "Review of life cycle assessment of energy payback and greenhouse gas emission of solar photovoltaic systems," Renewable and Sustainable Energy Reviews **19,** pp.255-274, 2013.
https://doi.org/10.1016/j.rser.2012.11.035

[11] First Solar Inc (2014), http://investor.firstsolar.com/releasedetail.cfm?ReleaseID=828273

[12] NREL, 2001. Available in www.nrel.gov/news/press/2001/1501_record.html

[13] J. Britt and C. Ferekides, "Thin film CdS/CdTe solar cell with 15.8% efficiency," Applied Physics Letters, **62(**22), p. 2851-2852, 1993.
https://doi.org/10.1063/1.109629

[14] Nowshad Amin, M A Matin, M M Aliyu, M A Alghoul, M R Karim and K Sopain, "Prospects of Back Surface Field Effect in Ultra-Thin High-Efficiency CdS/CdTe Solar Cells from Numerical Modeling," Int. J. Photoenergy, **2010**, pp. 1-8, 2010.
http://dx.doi.org/10.1155/2010/578580

[15] Minami, Tadatsugu, "Transparent conducting oxide semiconductors for transparent electrodes," Semiconductor Sci. and Technol. **20**(4), pp. S35-S44, 2005.
https://doi.org/10.1088/0268-1242/20/4/004

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 218-224 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Domain Independent Feature Extraction using Rule Based Approach

Sint Sint Aung[*,1], Myat Su Wai[2]

[1]*Department of Academic Affairs, University of Computer Studies, Mandalay, +95, Myanmar*

[2]*Web Mining Lab, University of Computer Studies, Mandalay, +95, Myanmar*

A R T I C L E I N F O

Article history:
Received: 16 November, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords:
Sentiment Analysis
Features Extraction
Opinions

A B S T R A C T

Sentiment analysis is one of the most popular information extraction tasks both from business and research prospective. From the standpoint of research, sentiment analysis relies on the methods developed for natural language processing and information extraction. One of the key aspects of it is the opinion word lexicon. Product's feature from online reviews is an important and challenging task in opinion mining. Opinion Mining or Sentiment Analysis is a Natural Language Processing and Information Extraction task that identifies the user's views or opinions. In this paper, we developed an approach to extract domain independent product features and opinions without using training examples i.e, lexicon-based approach. Noun phrases are extracted using not only dependency rules but also textblob noun phrase extraction tool. Dependency rules are predefined according to dependency patterns of words in the sentences. StandfordCoreNlp Dependency parser is used to identify the relations between words. The orientation of words is classified by using lexicon-based approach. According to the experimental results the system gets good performance in six different domains.

1. Introduction

Sentiment analysis is one of the most popular information extraction tasks both from business and research prospective. It has numerous business applications, such as evaluation of a product or company perception in social media [1]. From the standpoint of research, sentiment analysis relies on the methods developed for natural language processing and information extraction. One of the key aspects of it is the opinion word lexicon. Opinion words are such words that carry opinion. Positive words refer to some desired state, while negative words to some undesired one. For example, "good" and "beautiful" are positive opinion words, "bad" and "evil" are negative [2].

Opinion phrases and idioms exist as well. Many opinion words depend on context, like the word "large". Some opinion phrases are comparative rather than opinionated, for example "better than". Auxiliary words like negation can change sentiment orientation of a word[3].

[*]Corresponding Author: Sint Sint Aung, Email: ssaung@gmail.com

Opinion words are used in a number of sentiment analysis tasks. They include document and sentence sentiment classification, product features extraction, subjectivity detection etc. [4]. Opinion words are used as features in sentiment classification. Sentiment orientation of a product feature is usually computed based on the sentiment orientation of opinion words nearby [5]. Product features can be extracted with the help of phrase or dependency patterns that include opinion words and placeholders for product features themselves. Subjectivity detection highly relies on opinion word lists as well, because many opinionated phrases are subjective in [6]. Thus, opinion lexicon generation is an important sentiment analysis task. Detection of opinion word sentiment orientation is an accompanying task.

Opinion lexicon generation task can be solved in several ways. The authors of [7] point out three approaches: manual, dictionary-based and corpus-based. The manual approach is precise but time-consuming. The dictionary based approach relies on dictionaries such as WordNet. One starts from a small collection of opinion words and looks for their synonyms and antonyms in a dictionary [8]. The drawback of this approach is that the dictionary coverage

is limited and it is hard to create a domain-specific opinion word list. Corpus-based approaches rely on mining a review corpus and use methods employed in information extraction. The approach proposed in [9] is based on a seed list of opinion words. These words are used together with some linguistic constraints like "AND" or "OR" to mine additional opinion words.

Clustering is performed to label the mined words in the list as positive and negative. Part of speech patterns are used to populate the opinion word dictionary in and Internet search statistics is used to detect semantic orientation of a word. In [10], the authors extend the mentioned approaches and introduces a method for extraction of context-based opinion words together with their orientation. Classification techniques are used in [11] to filter out opinion words from text. The approaches described were applied in English. There are some works that deal with Russian. For example, in [12], the author proposed to use classification. Various features, such as word frequency, weirdness, and TF-IDF are used there.

Most of the research done in the field of sentiment analysis relies on the presence of annotated resources for a given language. However, there are methods which automatically generate resources for a target language, given that there are tools and resources available in the source language. Different approaches to multilingual subjectivity analysis are studied and are summarized in [13].

In one of them, subjectivity lexicon in the source language is translated with the use of a dictionary and employed for subjectivity classification. This approach delivers mediocre precision due to the use of the first translation option and due to word lemmatization. Another approach suggests translating the corpus. This can be done in three different ways: translating an annotated corpus in the source language and projecting its labels; automatic annotation of the corpus, translating it and projecting the labels; translating the corpus in the target language, automatic annotation of it and projecting the labels. Language Weaver 1 machine translation was used on English-Roman and English-Spanish data. Classification experiments with the produced corpora showed similar results. They are close to the case when test data is translated and annotated automatically. This shows that machine translation systems are good enough for translating opinionated datasets. In [14], the authors also confirmed when they used Google Translate 2, Microsoft Bing Translator 3 and Moses 4.

Multilingual opinion lexicon generation is considered in [that presents a semi-automatic approach with the use of triangulation. The authors use high-quality lexicons in two different languages and then translate them automatically into a third language with Google Translate. The words that are found in both translations are supposed to have good precision. It was proven for several languages including Russian with the manual check of the resulting lists. The same authors collect and examine entity-centered sentiment annotated parallel corpora [15].

The process of automatic extraction of knowledge by means of opinion of others about some particular product, topic or problem. Opinion mining is also called sentiment analysis due to large volume of opinion which is rich in web resources available online. Analyzing customer review is most important, it tend to rate the product and provide opinions for it which is been a challenging problem today. Opinion feature extraction is a sub problem of opinion mining, with the vast majority of existing work done in the product review domain. Main fields of research in sentiment analysis are Subjectivity Detection, Sentiment Prediction, Aspect based Sentiment Summarization, Text summarization for opinions, Contractive viewpoint, Summarization, Product Feature Extraction, Detecting opinion spam [16].

2. Related Works

Many researchers have addressed the problem of constructing subjective lexicon for different languages in recent years. In [17] to compile a subjective lexicon, the author investigated three main approaches and they are outlined in this section.

knowledge to extract the domain-specific sentiment lexicon based on constrained label propagation. According to [18], the authors had divided the whole strategy into six steps. Firstly, detected and extracted domain- specific sentiment terms by combining the chunk dependency parsing knowledge and prior generic sentiment lexicon. To refine the sentiment terms some filtering and pruning operations were carried out by others. Then they selected domain-independent sentiment seeds from the semi-structured domain reviews which had been designated manually or directly borrowed from other domains. As the third step, calculated the semantic associations between sentiment terms based on their distribution contexts in the domain corpus. For this calculation, the point-wise mutual information (PMI) was utilized which is commonly used in semantic linkage in information theory. Then, they defined and extracted some pair wise contextual and morphological constraints between sentiment terms to enhance the associations. The conjunctions like "and" and "as well as" were considered as the direct contextual constraints whereas "but" was referred to as a reverse contextual constraint. The above constraints propagated though out the entire collection of candidate sentiment terms. Finally, the propagated constraints were incorporated into label propagation for the construction of domain-specific sentiment lexicon. In [19], the authors proposed approach showed an accuracy increment of approximately 3% over the baseline methods. Opinion analysis has been studied by many researchers in recent years. Two main research directions are sentiment classification and feature-based opinion mining. Sentiment classification investigates ways to classify each review document as positive, negative, or neutral. Representative works on classification at the document level include. These works are different from ours as we are interested in opinions expressed on each product feature rather than the whole review.

In [20], ssentence level subjectivity classification is studied, which determines whether a sentence is a subjective sentence (but may not express a positive or negative opinion) or a factual one. Sentence level sentiment or opinion classification is studied in. Our work is different from the sentence level analysis as we identify opinions on each feature. A review sentence can contain multiple features, and the orientations of opinions expressed on the features can also be different, e.g., "the voice quality of this phone is great and so is the reception, but the battery life is short." "voice quality", "reception" and "battery life" are features. The opinion on "voice quality", "reception" are positive, and the opinion on

"battery life" is negative. Other related works at both the document and sentence levels include those in [21].

Most sentence level and even document level classification methods are based on identification of opinion words or phrases. There are basically two types of approaches: (1) corpus-based approaches, and (2) dictionary-based. approaches. Corpus-based approaches find co-occurrence patterns of words to determine the sentiments of words or phrases, e.g., the works in [22].

In [23], the authors proposed the idea of opinion mining and summarization. It uses a lexicon-based method to determine whether the opinion expressed on a product feature is positive or negative. In [24] and [25]. these methods are improved by a more sophisticated method based on relaxation labeling. We will show in Section 5 that the proposed technique performs much better than both these methods. In [26], a system is reported for analyzing movie reviews in the same framework. However, the system is domain specific. Other recent work related to sentiment analysis includes in. In [27], the authors studied the extraction of comparative sentences and relations, which is different from this work as we do not deal with comparative sentences in this research.

Our holistic lexicon-based approach to identifying the orientations of context dependent opinion words is closely related to works that identify domain opinion words . In [28], the authors used conjunction rules to find such words from large domain corpora. In [29], the conjunction rule basically states that when two opinion words are linked by "and" in a sentence, their opinion orientations are the same. For example, in the sentence, "this room is beautiful and spacious", both "beautiful" and "spacious" are positive opinion words. Based on this rule or language convention, if we do not know whether "spacious" is positive or negative, but know that "beautiful" is positive, we can infer that "spacious" is also positive. Although our approach will also use this linguistic rule or convention, our method is different in two aspects. First, we argue that finding domain opinion words is still problematic because in the same domain the same word may indicate different opinions depending on what features it is applied to. For example, in the following review sentences in the camera domain, "the battery life is very long" and "it takes a long time to focus", "long" is positive in the first sentence, but negative in the second. Thus, we need to consider both the feature and the opinion word rather than only the opinion word as in [30].

Opinion target and opinion word extraction are not new tasks in opinion mining. There is significant effort focused on these tasks. They can be divided into two categories: sentence-level extraction and corpus-level extraction according to their extraction aims. In sentence-level extraction, the task of opinion target/word extraction is to identify the opinion target mentions or opinion expressions in sentences. Thus, these tasks are usually regarded as sequence-labeling problems. Intuitively, contextual words are selected as the features to indicate opinion targets/words in sentences. Additionally, classical sequence labeling models are used to build the extractor, such as CRFs and HMM. In [31], the authors proposed a lexicalized HMM model to perform opinion mining. In [32], the authors used CRFs to extract opinion targets from reviews. However, these methods always need the labeled data to train the model. If the labeled training data are insufficient or come from the different domains than the current texts, they would have unsatisfied extraction performance. Although in [33], the authors proposed a method based on transfer learning to facilitate cross- domain extraction of opinion targets/words, their method still needed the labeled data from out-domains and the extraction performance heavily depended on the relevance between in-domain and out-domain.

Although many target extraction methods exist, we are not aware of any attempt to solve the proposed problem. According to [34], although in supervised target extraction, one can annotate entities and aspects with different labels, supervised methods need manually labeled training data, which is time-consuming and labor-intensive to produce. Note that relaxation labeling was used for sentiment classification in, but not for target classification.

3. Proposed method

This section presents the detailed of step by step process about the system.

3.1. Preprocessing the Input sentences

Input sentences with xml file are prepared before parsing to the StanfordCoreNLP parser. Xml tag are removed. ASCII code characters are replaced with Unicode characters because StanfordCoreNLP cannot process non-Unicode characters.

3.2. Rules for Features and Opinions Extraction

In this section, we describe how to extract opinion and product features using extraction rules. They are the most important tasks for text sentiment analysis, which has attracted much attention from many researchers. Based on the relations between features and opinions, there are four main rules in the double propagation;

1. extracting features using opinion words
2. extracting features using the extracted features
3. extracting opinion words using the extracted features
4. extracting opinion words using both the given and the extracted opinion words

In the following extraction rules, O is opinion word, H is the third word, {O} is a set of seed opinion lexicon, F is product feature, and O-Dep is part-of-speech information and dependency relations. {JJ}, {VB} and {NN} are sets of POS tags of potential opinion words and features, respectively. And {DR} contains dependency relations between features and opinions such as mod, pnmod, subj, s, obj, obj2, conj. We used rule 1 and 2 to extract features, and rule 3 and 4 use to extract opinion words. Moreover, we also used some additional patterns to extract features and opinions.

R11: If a word F whose POS is NN is directly depended by an opinion word O through one of the dependency relations mod, pnmod, subj, s, obj and obj2, then F is a feature. It can be defined as follows;

$$O \rightarrow O\text{-}Dep \rightarrow F$$

$$F \rightarrow F\text{-}Dep \rightarrow O$$

such that O ∈ {O}, O-Dep and F-Dep ∈ {DR}, where {DR} = {mod, pnmod, subj, s, obj, obj2, desc} and P OS (T) ∈ {N N}. For eg. Overall a sweet machine.

R12: If an opinion word O and a word F, whose POS is NN, directly depend on a third word H through dependency relations except conj, then F is a feature. It can be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$$

such that O ∈ {O}, O-Dep and F-Dep ∈ {DR}, POS(F) ∈ {NN}. For eg. Canon is the great product.

R13: If a word F whose POS is NN is indirectly depended by an opinion word O through another word H through two dependency relations dobj and amod or nmod:poss, then F is a feature. It can also be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow H \rightarrow F\text{-}Dep \rightarrow F$$

$$O \leftarrow O\text{-}Dep \leftarrow H \leftarrow F\text{-}Dep \leftarrow F$$

such that O ∈ {O}, O-Dep ∈ {DR}, F-Dep ∈ {DR}, POS(F) ∈ {NN}. For eg. I like the computer’ s battery.

R21: If a word Fj, whose POS is NN, directly depends on a feature Fi through conj, then Fj is a feature. It can also be expressed as follows;

$$Fi \rightarrow Fi - Dep \rightarrow Fj$$

such that Fi ∈ {F}, Fj-Dep ∈ {CONJ}, POS(Fj) ∈ {NN}. For eg. Overall, I like the system features and performance.

R22: If a word Fj, whose POS is NN, and a feature Fi, directly depend on a third word H through the same dependency relation, then Fj is a feature. It can also be expressed as follows;

$$Fi \rightarrow Fi\text{-}Dep \rightarrow H \leftarrow Fj\text{-}Dep \leftarrow Fj$$

such that Fi ∈ {F}, Fi-Dep ∈ {DR}, Fj-Dep ∈ {DR}, POS(Fj) ∈ {NN}. For eg. Canon has done an excellent job.

R31: If a word O whose POS is JJ or VB directly depends on a feature F through one of the dependency relations mod, pnmod, subj, s obj, obj2 and desc, then O is an opinion word. It can also be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow F$$

$$F \rightarrow F\text{-}Dep \rightarrow O$$

such that F ∈ {F}, O-Dep and F-Dep ∈ {DR}, POS(O) ∈ {JJ, VB}. For eg. Overall a sweet machine.

R32: If a word O whose POS is JJ or VB and a feature F directly depend on a third word H through dependency relations except conj, then O is an opinion word. It can also be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$$

such that F ∈ {F}, O-Dep and F-Dep ∈ {DR}, POS(O) ∈ {JJ, VB}. For eg. Canon is the great product.

R33: If a word O whose POS is JJ or VB indirectly depends on a feature F through another word H through dependency relations dobj, amod and nmod:poss, then O is an opinion word. It can also be expressed as follows;

$$O \rightarrow O\text{-}Dep \rightarrow H \rightarrow F\text{-}Dep \rightarrow F$$

$$O \leftarrow O\text{-}Dep \leftarrow H \leftarrow F\text{-}Dep \leftarrow F$$

such that F ∈ {F}, O-Dep and F-Dep ∈ {DR}, POS(O) ∈ {JJ}. For eg. I like the computer’ s battery.

R41: If a word Oj, whose POS is JJ or VB, directly depends on an opinion word Oi through dependency relation conj, then Oj is an opinion word. It can also be expressed as follows;

$$Oi \rightarrow Oi\text{-}Dep \rightarrow Oj$$

such that Oi ∈ {O}, Oi-Dep ∈ {CONJ}, POS(Oj) ∈ {JJ, VB}. For eg. Nice and compact.

R42: If a word Oj, whose POS is JJ or VB, and an opinion word Oi, directly depend on a third word H through the same dependency relation, then Oj is an opinion word. It can also be expressed as follows;

$$Oi \rightarrow Oi\text{-}Dep \rightarrow H \leftarrow Oj\text{-}Dep \leftarrow Oj$$

such that Oi ∈ {O}, Oi-Dep == Oj-Dep, POS(Oj) ∈ {JJ, VB}. For eg. The screen size and screen quality is amazing.

3.3. Features and Opinion Extraction

In this section This system takes raw data as input and xml tag and ascii code characters are removed. After that, word tokenization, part-of speech tagging and dependency identification between words are done by using StandfordfordCoreNLP dependency parser. We used the algorithm also from [17]. Table 1 shows some examples of English stop word list.

Table 1. Some English Stopwords List

to	Of	I	Me	My
Mine	You	At	They	In
Which	With	On	Under	Below
Above	Thing	Things	Some	Someone
Sometime	Something	Somebody	No one	nobody

To start the extraction process, a seed opinion lexicon, a list of general words, review data and extraction rules are input to the proposed algorithm. The extraction process uses a rule-based approach using the relations defined in above. The system assumed opinion words to be adjectives, adverbs and verbs in some cases. And product features are nouns or noun phrases and also verbs in some cases.

Its primary idea is that opinion words are usually associated with product features in some ways. Thus, opinion words can be recognized by identified features, and features can be identified by known opinion words. So, the extracted opinion words and product features are used to identify new opinion words and new product features. The extraction process ends when no more opinion words or product features can be found.

Table 2. Some Unigram Features

Price	System	Computer	Laptop	Reviews
Customer	Battery	Camera	Power	Screen
Staff	Food	Wine	Window	Notebook
Work	Use	Download	Connection	Cost
D-link	Device	Feature	Model	quality

Moreover, textblob is used to extract ngram noun phrase words from the sentences that are not covered with extraction rules. It can increase the performance of the system in term of precision, recall, and f1-score. In order to increase the accuracy, stop words and some general words are removed during the extraction time. Table 2, 3 and 4 describe some extracted of unigrams, bigrams, trigrams and n-gram words from the system.

Table 3. Some Bigram Features

Picture quality	Battery life	Tech support	Screen quality
Power supply	Printer sharing	Return policy	Set up
Security setting	Service tech	Setup software	Mac support
Guest feature	8GB RAM	web cam	charger unit
connect quality	cooling system	cordless mouse	data rate
Window 7	Cd drive	Hard drive	Service tech

The system constructs n-gram dictionary to refine the noun phrase extraction. If the phrase contains in this dictionary, the system extracts it as a feature. Otherwise, remove it.

Table 4. Sme N-gram Features

D-Link support crew	customer service agents
cover for the DVD drive	Dell's customer disservice
D-Link support crew	extended life battery
Garmin GPS software	sound quality via USB
customer service center	design based programs
direct Electrical connectivity	built it web cam
fingerprint reader driver	technical service for dell

4. Classifying Polarity Orientation

Opinions are classified by using Vader lexicon and qualitative analysis techniques developed by C.J. Hutto and Eric Gilbert, 2014. This deep qualitative analysis resulted in isolating five generalizable heuristics based on grammatical and syntactical cues to convey changes to sentiment intensity. They incorporate word-order sensitive relationships between terms:

Punctuation, namely the exclamation point (!), increases the magnitude of the intensity without modifying the semantic orientation.

Capitalization, specifically using ALL-CAPS to emphasize a sentiment relevant word in the presence of other non-capitalized words, increases the magnitude of the sentiment intensity without affecting the semantic orientation

Degree modifiers (also called intensifiers, booster words, or degree adverbs) impact sentiment intensity by either increasing or decreasing the intensity.

The **contrastive conjunction** "but" signals a shift in sentiment polarity with the sentiment.

By examining the **tri-gram** to deeply analyze the intensity of sentiment orientation.

Table 5. shows some example of polarity classification. Polarity score included negative sign (-) indicates negative opinion. And, score with positive sign (+) indicates positive opinion.

Table 5. Polarity Classification of the System

Words	Polarity Score	Polarity Label
Set up	0.4404	positive
excellent	0.5719	positive
work	0.7264	positive
installation disk	-0.296	negative
problem	-0.4019	negative

5. Experimental Results

For experiment, we use core i7 processor, 4GB RAM and 64-bit Ubuntu OS, And, we implement the proposed system with python programming language (PyCharm 2016.3 IDE for python).

Table 6. Dataset Used in the System

Dataset	no of sentences	no of features
ABSA15 restaurant	1083	1193
ABSA15 Hotel	266	212
Router	245	304
Speaker	291	435
Computer	239	346
iPod	161	293
Linksys Router	192	375
Nokia 6000	363	633
Norton	210	302
Diaper Champ	212	239

In this paper, 10 product review datasets are collected and evaluated in term of precision, recall and f1-score. Table 6 shows the domains according to their names, the number of sentences

and the number of features. For performance evaluation on product feature extraction, the comparative results between the proposed approach and Qiu's approach are also analyzed. Precision, recall and f1-score of both are described in figure 1, 2 and 3.

From figure 1, we can see that the proposed approach has higher precision in 6 datasets, dropped in 3 datasets and nearly the same in 1 dataset over Qiu's approach. The proposed approach relies only on the review data itself and no external information is needed.

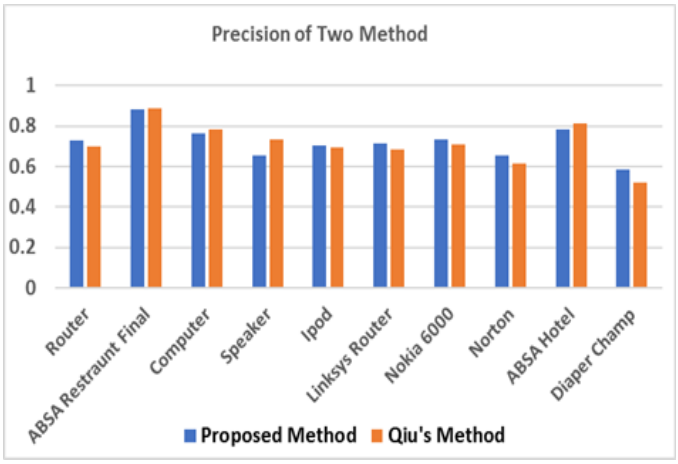


Figure 1: Comparison of Precision on Feature Extraction between Two Approaches

According to the experimental results, the highest precision 0.8819 (88%) are achieved in ABSA Restaurant dataset. In this dataset, the precision, recall and f1-score are nearly the same because there are no verb product features in this dataset.

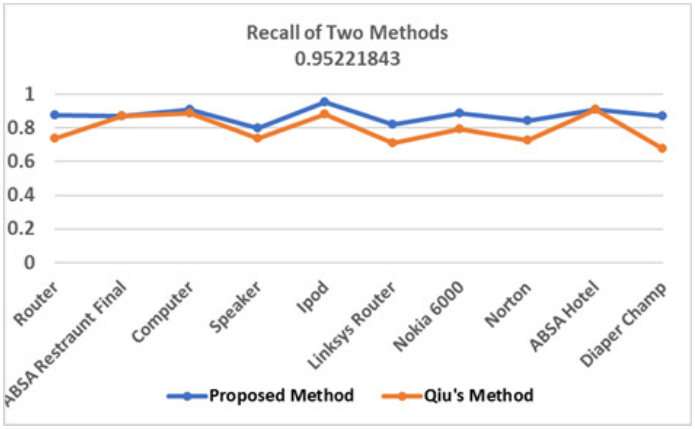


Figure 2: Comparison of Recall on Feature Extraction between Two Approaches

Figure 2 shows that proposed approach outperforms all the other approaches in recall except ABSA Restaurant dataset and ABSA Hotel dataset. The proposed approach has about 14% improvement in Router dataset, about 3% in Computer dataset, about 6% in Speaker dataset, about 7% in iPod dataset, about 10% in Linksys Router dataset, about 9% in Nokia 6000 dataset, about 12% in Norton dataset, about 20% in Diaper Champ dataset. In ABSA Restaurant and ABSA Hotel datasets, the recall of the two approaches are the same. So, to sum up, the proposed approach outperforms over the Qiu's approach according to the recall.

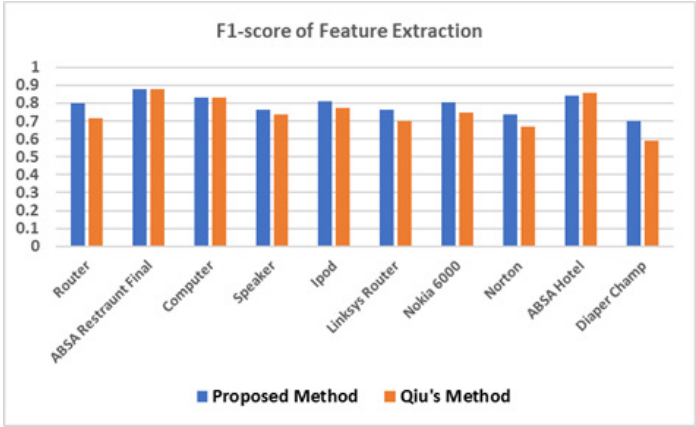


Figure 3: Comparison of F1-score on Feature Extraction between Two Approaches

The comparative results of f1-score in feature extraction are shown in figure 3. According to the experimental results, the proposed approach has higher f1-score (about 7%) in 7 datasets and about 0.003% drop in ABSA Restaurant and Computer datasets and 1% drop in ABSA Hotel dataset

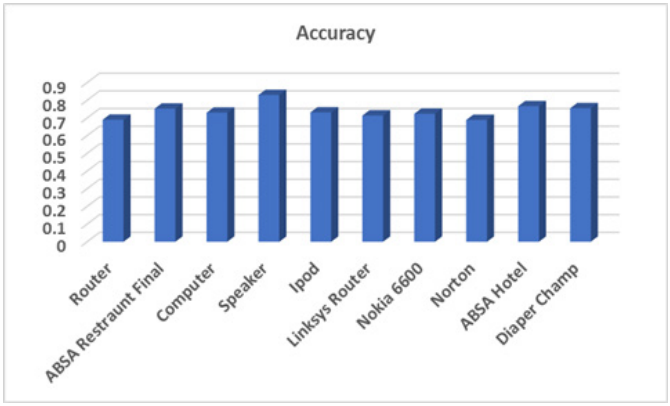


Figure 4 Experimental results of Polarity Classification

The performance analysis of polarity classification of the proposed system is evaluated in all datasets. Figure 4 shows the experimental results of polarity classification in all datasets. The system achieves Highest accuracy 83% in speaker dataset.

6. Conclusion

Opinion mining and sentiment classification are not only technically challenging because of the need for natural language processing. In this work, an effective opinion lexicon expansion and feature extraction approach is proposed. Features and opinions words are extracted simultaneously by using proposed algorithm based on double propagation. So, unlike the existing approach, context dependent opinion words are extracted and domain independence. According to experimental results, the proposed system works well in all datasets and get domain independency without using training examples. As the future extension, we will analyze the performance of the proposed system with more different datasets from SemEval research group.

And we will apply more dependency relations in extraction process.

References

[1] Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., and Hussain, A. "Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach". Cognitive Computation, pages 1–13, 2015.

[2] Asghar M, Khan A, Ahmad S, Kundi F, "A Review of Feature Extraction in Sentiment Analysis", Journal of Basic and Applied Scientific Research, 4(3): 181–186, 2014.

[3] Broß, J. "Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques". PhD thesis, Freie Universität Berlin, 2013.

[4] Chen, L., Wang, W., Nagarajan, M., Wang, S., Sheth, A. P. "Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter", In the Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media(ICWSM), 50–57, 2012.

[5] Chinsha, T. and Joseph, S. "A syntactic approach for aspect-based opinion mining". In Semantic Computing (ICSC), 2015 IEEE International Conference on, pages 24–31. IEEE.

[6] De Marneffe, M.-C. and Manning, C. D. "Stanford typed dependencies manual. Report", Technical report, Stanford University, 2016.

[7] Fabbrizio G, Aker A, Gaizauskas R012) "Summarizing Online Reviews Using Aspect Rating Distributions and Language Modeling", IEEE Intelligent Systems, 28(3): 28–37, 2012.

[8] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," Computational Linguistics 37, no.1, 2011, pp. 9–27.

[9] Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," In Proc. of the 49th Ann. Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 151-160. Association for Computational Linguistics, 2011.

[10] Khan, K., Baharudin, B., Khan, A., and Ullah, A. "Mining opinion components from unstructured reviews: A review". Journal of King Saud University-Computer and Information Sciences, 26(3):258–275, 2014.

[11] Kundi F, Ahmad S, Khan A, Asghar, "Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet", Life Science Journal, 11(9):66–72, 2014.

[12] Liu, B. "Web data mining: exploring hyperlinks, contents, and usage data". Springer Science & Business Media, 2011.

[13] Liu, B. "Sentiment analysis and opinion mining". Synthesis Lectures on Human Language Technologies, 5(1):1–167, 2015.

[14] Liu, B. "Sentiment analysis mining opinions, sentiments, and emotions". 1:1–386, 2015.

[15] Liu, L. Xu and J. Zhao, "Opinion target extraction using word-based translation model", Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1346-1356, 2012.

[16] Moghaddam, A. S., "Aspect-based opinion mining in online reviews". PhD thesis, Applied Sciences: School of Computing Science, 2014.

[17] Myat Su Wai and Sint Sint Aung, "Simultaneous Opinion Lexicon Expansion and Product Feature Extraction", Proceeding of 16th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2017), ISBN: 978–1–5090–5506–7.

[18] Neviarouskaya A, Aono M, "Sentiment Word Relations with Affect, Judgment, and Appreciation", IEEE Transactions on Affective Computing, 4(4): 425–438, 2014.

[19] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I., "Semeval-2015 Task 12: Aspect Based Sentiment Analysis". In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 486–495, Denver, Colorado. Association for Computational Linguistics.

[20] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S.. "Semeval-2014 Task 4: Aspect Based Sentiment Analysis". In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

[21] S. Poria, B. Agarwal, Gelbukh, A. Hussain, N. Howard, "Dependency-based semantic parsing for concept-level text analysis", In Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Proceeding, Part I. LNCS, vol. 8403, pp. 113–127. Springer, Heidelberg, 2014.

[22] S. Poria, E. Cambria, G. Winterstein, G. B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis,". Knowledge-Based Systems, vol.69, pp. 45-63, 2014.

[23] S. Poria, E. Cambria, L.W. Ku, C. Gui, A. Gelbukh, "A rule-based approach to aspect extraction from product reviews," In Workshop Proc. of the 25th International Conference on Computational Linguistics, COLING'14, pp. 28-37, 2014.

[24] Shariaty, S. and Moghaddam, S.. "Fine-grained opinion mining using conditional random fields". In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pages 109–114. IEEE.

[25] Souza M, Vieira R, Busetti D, Chishman R, Alves I, "Construction of a Portuguese Opinion Lexicon from Multiple Resources", in the proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL' 2011), 59–66. Springer, pp.411-448, 2007.

[26] Xu, T. J. Zhao, D. Q. Zheng, S. Y. Wang, "Product features mining based on Conditional Random Fields model", Proceedings of the 2010 International Conference on Machine Learning and Cybernetics, pp. 3353-3357, 2010.

[27] Xu, X., Cheng, X., Tan, S., Liu, Y., and Shen, H. "Aspect-level opinion mining of online reviews". China Communications, 10(3):25–41.

[28] Y. Wu, Q. Zhang, X. Huang, and L. Wu, "Phrase dependency parsing for opinion mining,". In EMNLP'09, Volume 3-Volume 3, pp. 1533- 1541. Association for Computational Linguistics, 2009.

[29] Y. Zheng, L. Ye, G. Wu and X. Li, "Extracting product features from Chinese customer reviews," Proceedings of 3rd International Conference on Intelligent System and Knowledge Engineering, pp. 285-290, 2008.

[30] Z. Hai, K. Chang, G. Cong, "One seed to find them all: mining opinion features via association", Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 255-264, 2012

[31] Zhai, Z., Liu, B., Xu, H., and Jia, P., "Clustering product features for opinion mining". In Proceedings of the fourth ACM international conference on Web search and data mining, pages 347–354. ACM.

[32] Zhang D, Dong H, Yi J, Song L, "Opinion summarization of customer reviews", In proceedings of the International Conference on Automatic Control and Artificial Intelligence (ACAI 2012), 1476–1479.

[33] Zhang, L. and Liu, B. "Aspect and entity extraction for opinion mining". In Data mining and knowledge discovery for big data, pages 1–40. Springer.

[34] Zhang, Y. and Zhu, W. "Extracting implicit features in online customer reviews", In Proceedings of the 15th ACM international conference on Information and knowledge management, pages 43–50. ACM.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 225-233 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Velocity obstacles for car-like mobile robots: Determination of colliding velocity and curvature pairs

Emese Gincsainé Szádeczky-Kardoss[*,1], Zoltán Gyenes[2]

[1]*Associate professor at Department of Control Engineering and Information Technology, Budapest University of Technology and Economics, H-1117, Hungary*

[2]*Student at Department of Control Engineering and Information Technology, Budapest University of Technology and Economics, H-1117, Hungary*

ARTICLE INFO

Article history:
Received: 29 November, 2017
Accepted: 15 January, 2018
Online: 30 January, 2018

Keywords:
Mobile robots
Motion planning
Velocity obstacles

ABSTRACT

This paper addresses the motion planning problem of Reeds-Shepp-type car-like mobile robots moving among static and dynamic obstacles. If the positions and the velocity vectors of the obstacles are known or well estimated, the Velocity Obstacles (VO) method and its non-linear version (NLVO) can be used to plan a collision-free trajectory for a robot in the dynamic environment. VO and NLVO algorithms determine a velocity vector for the robot which corresponds not necessarily to the orientation of the robot, hence a nonholonomic car-like mobile robot cannot apply it exactly. Previously, the NLVO method was adopted for Dubins-like mobile robots, which are able to move forward only. In this paper, a method similar to NLVO is presented, but it results motions feasible for Reeds-Shepp-type robots, which are able to drive both forward and backward. Longitudinal velocities and curvatures of turning circles are calculated, which ensure collision-free motion if the arbitrary movement of the obstacles are known for some time-horizon.

1. Introduction

This paper is an extension of work originally presented in the 25th Mediterranean Conference on Control and Automation [1].

One of the main tasks of autonomous mobile robots is to execute a safe motion in their workspace from the actual position to a desired goal configuration. In some applications, the environment is fix, i.e. the positions and velocities of the obstacles are known. Otherwise the robot has to use some sensors to be able to estimate this information. Robots should be able to plan their motions such that they will not collide with static or moving obstacles. The literature presents several path planning methods for the avoidance of static obstacles (see e.g. [2,3]).

There are two possibilities for motion planning among moving obstacles: 1, The planning can be done in two steps: geometric path planning and then velocity planning. 2, The geometry of the path and the velocity profile along it are determined simultaneously.

[*]Corresponding Author: Emese Gincsainé Szádeczky-Kardoss,
Email: szadeczky@iit.bme.hu

In the first case, the geometry of the path and its time distribution are calculated separately (e.g. [4,5]). At the path planning phase, the moving obstacles are disregarded. The planner calculates a path which ensures collision-free motion to the goal among the static obstacles. The motions of the obstacles influence only the velocity profile of the path.

The second possibility is to calculate both the path geometry and the velocity profile in one step (e.g. [6]). In this second case the shape of the path depends also on the positions and movements of the dynamic obstacles.

If the information about the obstacles (positions, velocities) are known, a global path planner can be applied. If the environment is unknown, a local reactive obstacle avoidance algorithm has to be used based on the sensory information (e.g. [7]).

Several motion planning methods in static environment can be extended to solve the dynamic problem as well. Some methods use the configuration space of the robot to find a feasible path among static obstacles (e.g. [3]). If the workspace of the robot contains moving obstacles as well, the configuration space should be

extended by a temporal dimension, and the motion should be searched in this extended space. Using this solution, one has to modify the distance metric to deal with the temporal dimension [5].

The Artificial Potential Field (APF) algorithm is a quite simple method which can be used for path planning in static environment [8]. Applying special potential force functions, one can solve the motion planning problem in case of moving target and obstacles using APF [9,10].

The Dynamic Window (DW) method is a local obstacle avoidance algorithm used in static environment [11]. The planning is done in the velocity space of the robot. Reachable and admissible velocity values are selected. Reachable values satisfy the kinematic and dynamic constraints of the robot and admissible values guarantee that the robot can stop before hitting an obstacle. An adaptation of DW can be used with moving obstacles as well [7].

The inevitable collision states (ICS) for robots are presented in [12]. If the robot is in an IC state, it surely collides with an obstacle independently from the future trajectory of the robot. If a state is non-ICS, there exists at least one motion possibility for the robot to avoid collision. Using the ICS concept, the motion planning problem in dynamic environment can also be solved [13].

The concept of velocity obstacles (VO) was introduced in [6] for such environments where the velocity vectors of the obstacles are supposed to be unchanged for some time-interval. Using VOs, an avoidance maneuver can be determined in the velocity space of the robot, based on the current positions and velocities of the robot and obstacles. Velocity obstacles represent the set of robot velocities that would result in a collision with a static or moving obstacle. The basic VO method was extended for obstacles moving along arbitrary trajectories (non-linear velocity obstacles – NLVO [14]).

The inverse version of NLVO (INLVO) can be used to plan the velocity profile for a robot along a path with fix geometry [15]. A modified VO method can be used to plan autonomous navigation for unmanned surface vehicles as well [16]. The hybrid reciprocal VO (HRVO) is a method to plan the motion of multiple mobile robots without central coordination [17].

Our conference paper [1] presented a modified version of VO to plan the motion for Dubins-like mobile robots (VOD - Velocity Obstacles for Dubins-like robots). These robots go only forward. VOD defined pairs of colliding velocity and turning radius.

In this paper, the goal is to define velocity obstacles for Reeds-Shepp-type car-like mobile robots, which are able to drive both forward and backward. The determination of colliding velocity and curvature pairs is discussed. In this paper, the curvature is used instead of turning radius, hence the graphical representation of the colliding pairs is easier. The method and the equations of [1] were modified to deal with negative velocities and to use curvature instead of turning radius.

The paper is organized as follows. Section 2 gives a short review of velocity obstacle methods. Section 3 presents the properties of Reeds-Shepp-type mobile robots. Section 4 describes how the 'velocity' of a car-like robot is represented in this work. In Section 5, the velocity obstacles for car-like mobile robots (VOCL) are presented. Some simulation results are given in Section 6. Finally, a short section concludes the paper.

2. A Short Review of Velocity Obstacles

VO method can be used to find a feasible velocity vector for the robot such that the robot is able to avoid static and moving obstacles. It is assumed that the position and velocities of the obstacles are known, and the obstacles follow a straight-line path for some time-horizon. The VO method uses circular representation of the robot and the obstacles with known radii.

Given are a robot A and some moving obstacles B_i ($i = 1 \dots n < \infty$), where n denotes the number of obstacles. (According to this concept, a static obstacle is a moving obstacle with zero velocity.) The velocity obstacle VO_i contains all the robot velocity vectors $\mathbf{v}$ which would result a collision with obstacle B_i:

$$VO_i = \{\mathbf{v} | \exists t: A(t, \mathbf{v}) \cap B_i(t) \neq \emptyset\} \quad (1)$$

where $A(t, \mathbf{v})$ denotes the robot at time t if velocity $\mathbf{v}$ was applied. The shape of VO_i is a cone (see Figure *1*). The union of the individual VO_i reads

$$VO = \cup_{i=1}^{n} VO_i. \quad (2)$$

Selecting a velocity vector outside VO guarantees that no collision will occur between the robot and the obstacles.

An example for a point robot A and two moving (B_1, B_2) and a static obstacle (B_3) is presented in Figure *1*.

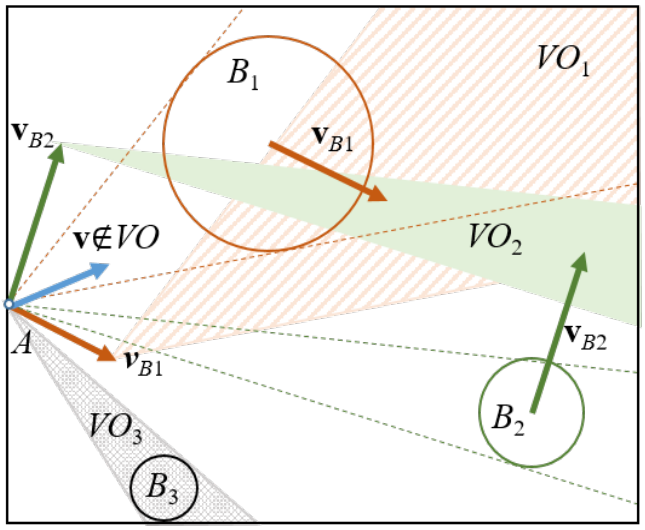


Figure 1. Velocity obstacles for a point robot A with two moving (B_1, B_2) and a static obstacle (B_3). $\boldsymbol{v}_{B1}, \boldsymbol{v}_{B2}$ are the velocities of the obstacles. VO_i denotes the velocity obstacle corresponding to obstacle B_i. A collision-free example is depicted for robot velocity: $\boldsymbol{v} \notin VO = \cup_{i=1}^{3} VO_i$.

2.1. Non-Linear Velocity Obstacles

The non-linear velocity obstacle (NLVO) defines the set of all linear robot velocities that would result a collision with obstacle B_i (t) moving along arbitrary known trajectory. At time instant t,

one can define robot velocity vectors which move the robot to a position during time $t - t_0$ such that a collision occurs with B_i:

$$NLVO_i(t) = \frac{c_i(t)+B_i}{t-t_0} \tag{3}$$

where $c_i(t)$ denotes the trajectory of the obstacle B_i. Considering all $t > t_0$, one gets

$$NLVO_i = \bigcup_{t>t_0} \frac{c_i(t)+B_i}{t-t_0}. \tag{4}$$

The shape of the non-linear velocity obstacle $NLVO_i$ is a warped cone with apex at A (see Figure 2). Similar to (2), the union of the individual $NLVO_i$ defines the set of all robot velocity vectors which result a collision for some $t > t_0$.

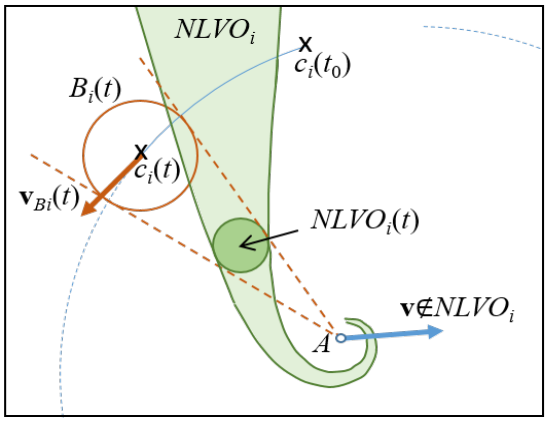


Figure 2. $NLVO$ for a point robot A with obstacle B_i moving along a circular path.

2.2. Generalized Velocity Obstacles

The concept of velocity obstacles was generalized to apply it for car-like robots [18]. The obstacle is defined in the control space of the robot:

$$GVO = \{u | \exists t: \|A(t,u) - B(t)\| < r_A + r_B\} \tag{5}$$

where u is the control input of the robot, $A(t,u)$ is the position of A if control u was applied up to t. r_A and r_B denote the radii of the circular robot and obstacle.

The controls are sampled, and for each control u, the minimum distance between A and B is determined numerically, if control u was applied to the robot. If the minimum distance is smaller than the sum of the radii, $u \in GVO$ and it means that a collision will occur between B and A for the control u.

2.3. Idea for Extension

In this paper, the velocity obstacles are applied to similar robots as presented in Subsection 2.2. This solution will not be restricted to sampled control inputs, as suggested by [18]. The presented method represents the velocity obstacles for car-like robots as a subset of a two-dimensional plane similar to the methods of VO and NLVO. In this work, this plane is determined by the velocity of the robot and the curvature of its path.

3. Reeds-Shepp-Type Car-Like Mobile Robots

A car-like mobile robot (see Figure 3) can move in a two-dimensional workspace. The state of the robot is defined by its position and orientation. The position is given by the (x, y) coordinates of the midpoint of its rear axle. The orientation of the robot is denoted by θ, and it is defined by the angle of the positive x axis of the coordinate system and the longitudinal axis of the robot. Ackermann-steering is supposed (see [19]), and the movement of the robot is described by the motion of a bicycle putting on the longitudinal axis of the robot. The inputs of the robot are the longitudinal velocity v and the angle of the front turning wheel (δ). Notice, that v is a scalar.

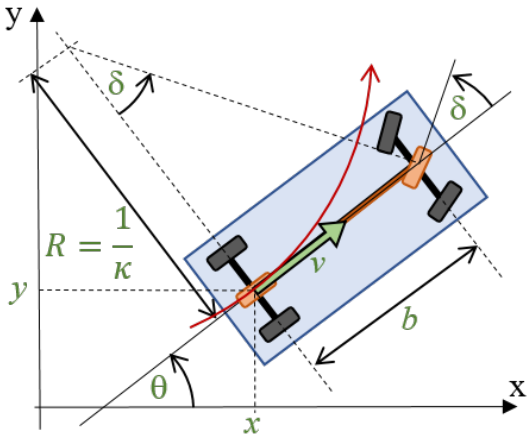


Figure 3. Car-like mobile robot.

The kinematics of the robot reads:

$$\begin{bmatrix}\dot{x}\\ \dot{y}\\ \dot{\theta}\end{bmatrix} = \begin{bmatrix}\cos\theta\\ \sin\theta\\ \frac{\tan\delta}{b}\end{bmatrix} v, \tag{6}$$

where b is distance between the front and the rear wheels. The turning angle δ of the front wheel determines the radius R of the turning circle of the robot: $R = \frac{b}{\tan\delta}$. The reciprocal of the turning radius is called curvature: $\kappa = \frac{1}{R} = \frac{\tan\delta}{b}$.

The turning radius R cannot be arbitrary small, since the turning angle δ of the front wheels is also limited. Depending on the limit of δ and the geometrical b parameter of the robot, a minimal turning radius can be determined: $R_{min} \leq |R|$. Similarly, the curvature is also limited: $|\kappa| \leq \kappa_{max}$.

The car-like mobile robot is a nonholonomic system since the direction of its velocity vector is constrained by the following equation:

$$\dot{x}\sin\theta - \dot{y}\cos\theta = 0. \tag{7}$$

There are different types of car-like mobile robots: Dubins-robots can only move forward (i.e. $v > 0$). It was proved that for given start and goal states the path with the minimal length consists of three segments: two or three circular segments with maximal

curvature (i.e. with minimal turning radius) and, if needed, a straight-line between two circles if no obstacles are presented in the workspace [20]. The Reeds-Shepp-type mobile robots are able to travel both forward and backward. It was also shown that the path with minimal length between given start and goal states contains (like Dubins-robots) circular segments with maximal curvature, and it may contain straight-lines and cusp points where the driving direction changes [21].

4. Velocity Representation for Car-Like Mobile Robots

A Reeds-Shepp-type mobile robot can move forward or backward on a circular path segment with constant curvature (κ), i.e. with constant radius ($R = \frac{1}{\kappa}$), or on a straight-line. The curvature may have negative values, which shows that the robot turns in the negative direction (i.e. clockwise). The straight-line motion can also be described by a zero curvature. The time-distribution along the path depends on the longitudinal velocity v of the robot. The sign of v shows the direction of motion. These imply that the actual motion of the robot can be described by the pair (v, κ).

If (v, κ) is constant for a time period Δt, one can calculate the displacement in position and in orientation for Δt, if the motion starts at $t_0 = 0$ from $[x_0, y_0, \theta_0] = [0,0,0]$ by integrating (6):

$$\begin{bmatrix} x(\Delta t) \\ y(\Delta t) \\ \theta(\Delta t) \end{bmatrix} = \begin{bmatrix} \frac{2}{\kappa} \sin\frac{\kappa v \Delta t}{2} \cos\frac{\kappa v \Delta t}{2} \\ \frac{2}{\kappa} \sin^2\frac{\kappa v \Delta t}{2} \\ \kappa v \Delta t \end{bmatrix}. \quad (8)$$

This works in the reverse direction as well: if a point (x, y) is given in the workspace of the robot, one can determine the constant curvature κ and the constant forward longitudinal velocity $v^f \geq 0$ and backward velocity $v^b < 0$, which can move the robot from $t_0 = 0$ and $[x_0, y_0, \theta_0] = [0,0,0]$ to this point during time Δt:

$$\kappa = \frac{2y}{x^2+y^2}, \quad (9)$$

$$v^f = \frac{x^2+y^2}{y\Delta t} \operatorname{atan}\frac{y}{x}, \quad (10)$$

$$v^b = \frac{x^2+y^2}{y\Delta t} \left(\operatorname{atan}\frac{y}{x} - sgn\left(\operatorname{atan}\frac{y}{x}\right)\pi\right). \quad (11)$$

Notice, that (10) (and (11)) has several solutions for given x and y. If the robot does not go more than ones around in a circle, (10) (or (11)) will have a single solution for v^f (or for v^b) with $-\pi < \operatorname{atan}\frac{y}{x} \leq \pi$.

If the robot goes straight, its motion is described by: $x = v\,\Delta t, y = 0, \kappa = 0, R = \infty, v = \frac{x}{\Delta t}$. In the sequel, the goal is to determine the (v, κ) (i.e. (v^f, κ) and (v^b, κ)) pairs which result a collision-free motion for the robot.

5. Velocity Obstacles for Car-Like Robots

Let the robot be at the $[x_0, y_0, \theta_0] = [0,0,0]$ initial state. First, such (v, κ) pairs are determined which result a collision with static obstacles. Then the collision with moving obstacles is also considered.

5.1. Collision with Static Obstacles

Suppose that a static obstacle is presented in the workspace of the robot. Let a circle represent the obstacle. Its position is given by the coordinates of its center: (x_o, y_o) and its radius is r_o. For the sake of simplicity, the robot is also represented by a circle with radius r_r. (A non-circular robot or obstacle can be approximated by one or more circles.) To examine the collision for a point-like robot, the radius of the obstacle should be enlarged by the radius of the robot: $r_o' = r_o + r_r$ similar to the method proposed in [6].

First, such circular motions (i.e. curvatures κ_g) are considered, where the robot grazes (touches) the obstacle. Geometrically, the problem is the following: given a point (position of the robot), a straight-line going through this point (line of the robot's orientation) and a circle (obstacle with enlarged radius r_o'), such a circle has to be found, which is tangential to the circle of the obstacle and grazes the line of orientation at the given point (robot's position).

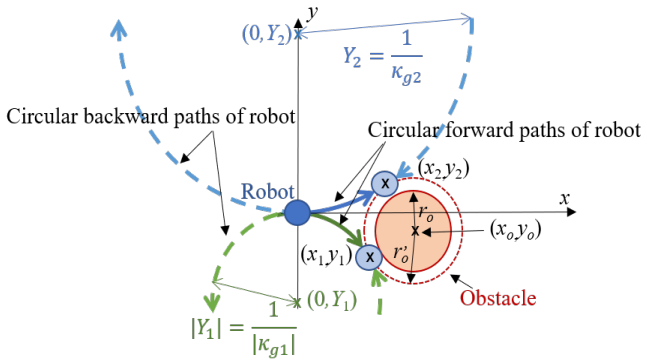


Figure 4. Grazing a static obstacle.

There are two possibilities (see Figure 4). In both cases the center of the turning circle lies on the y-axis (i.e. the x coordinate of the center of the turning circle is 0). The y coordinate of the center of the turning circle defines the radius of the circle. The two possible values for the curvatures (or for reciprocals of y coordinates of the center) are:

$$\kappa_{g1} = \frac{1}{Y_1} = \frac{2(y_o - r_o')}{x_o^2 + y_o^2 - r_o'^2}, \quad (12)$$

$$\kappa_{g2} = \frac{1}{Y_2} = \frac{2(y_o + r_o')}{x_o^2 + y_o^2 - r_o'^2}. \quad (13)$$

All curvatures between κ_{g1} and κ_{g2} will result in a collision with the static obstacle. More precisely, moving on a circle with curvature $\kappa_i \neq 0$ will cause a collision if

$$\min(\kappa_{g1}, \kappa_{g2}) < \kappa_i < \max(\kappa_{g1}, \kappa_{g2}). \quad (14)$$

$\kappa_i = 0$ can only cause a collision if

$$sgn(\kappa_{g1}) \neq sgn(\kappa_{g2}) \wedge sgn(x_o) = sgn(v). \quad (15)$$

The coordinates of the grazing points ((x_1, y_1) resp. (x_2, y_2)) can be calculated as well:

$$x_1 = \frac{x_o(x_o^2+y_o^2-r_o'^2)}{x_o^2+(y_o-r_o')^2}, \quad (16)$$

$$y_1 = \frac{(x_o^2+y_o^2-r_o'^2)(y_o-r_o')}{x_o^2+(y_o-r_o')^2}, \quad (17)$$

$$x_2 = \frac{x_o(x_o^2+y_o^2-r_o'^2)}{x_o^2+(y_o+r_o')^2}, \quad (18)$$

$$y_2 = \frac{(x_o^2+y_o^2-r_o'^2)(y_o+r_o')}{x_o^2+(y_o+r_o')^2}. \quad (19)$$

Consider now the time instant t. The (v, κ) input pairs can be determined, which cause collision with the static obstacle B_i at t. The velocity obstacle for a car-like robot (VOCL) is the union of these points (see Figure 5):

$$VOCL_i(t) = \{ (v, \kappa) | A(t, v, \kappa) \cap B_i \neq \emptyset \}, \quad (20)$$

where $A(t, v, \kappa)$ represents the robot at time-moment t moving from $t_0 = 0$ and from the origin on a circular path with radius $\frac{1}{\kappa}$ with velocity v. κ and v can be calculated from (x, y) position using (9)-(11).

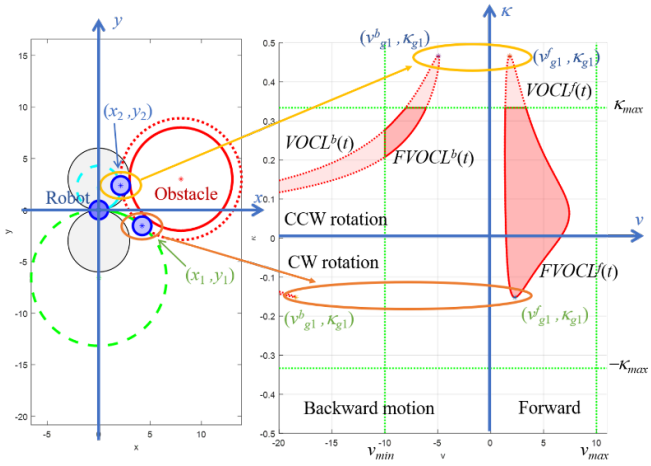


Figure 5. a. Robot and obstacle at a grazing case. b. Velocity obstacle $VOCL(t)$ and feasible VO $FVOCL(t)$ for time instant t

From the grazing points (x_1, y_1) and (x_2, y_2), one can determine the corresponding points in $VOCL_i(t)$. The curvatures κ_{g1} and κ_{g2} are used and the forward velocities v_{g1}^f, v_{g2}^f and backward velocities v_{g1}^b, v_{g2}^b can also be calculated if (x_1, y_1) or (x_2, y_2) should be reached on a circular path during time t.

(v_{g1}, κ_{g1}) and (v_{g2}, κ_{g2}) lie on the boundary of $VOCL_i(t)$. All other points on the boundary of $VOCL_i(t)$ correspond to a motion (i.e. (v, κ) values) which moves the robot during time t to a boundary point of the obstacle, but before or after the time instant t a collision occurs.

$VOCL_i(t)$ according to (20) may contain (v_j, κ_j) pairs, such that $|\kappa_j| > \kappa_{max}$ or $|v_j| > v_{max}$. These κ_j or v_j values are not feasible. Feasible VO ($FVOCL_i(t) \subseteq VOCL_i(t)$) contains only feasible curvature and velocity values:

$$FVOCL_i(t) = \{(v, \kappa) | A(t, v, \kappa) \cap B_i \neq \emptyset \wedge$$

$$|v| \leq v_{max} \wedge |\kappa| \leq \kappa_{max}\}. \quad (21)$$

In Figure 5 $VOCL_i(t)$ is delimited by dotted line and $FVOCL_i(t)$ is bounded by solid line.

Taking time-moments in a time-interval $t \in [t_0, t_h]$ one can get $FVOCL_i$ for a static obstacle B_i:

$$FVOCL_i = \bigcup_{t \in [t_0, t_h]} FVOCL_i(t). \quad (22)$$

If the robot moves according to $(v, \kappa) \notin FVOCL_i$, the robot will not collide with the static obstacle B_i in the given time interval $[t_0, t_h]$.

Notice, that the selection of t_h influences the effectiveness and the computational demand of the method. If a small value was selected for t_h, a mobile robot may collide with a moving obstacle. On the other hand, a large time-horizon can result that VO includes almost the complete velocity space [13].

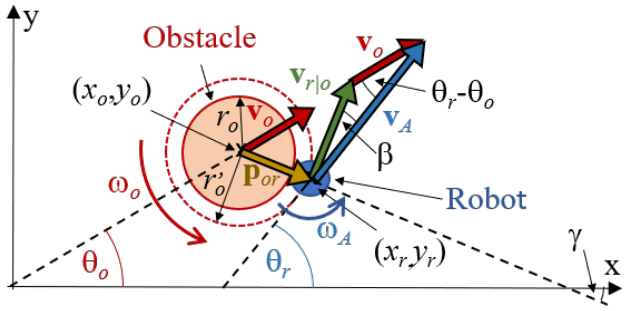
5.2. Calculating VOCL for Moving Obstacles

For a moving obstacle $B_i(t)$ the definition of $VOCL_i(t)$ for a given time-moment t is similar to (20):

$$VOCL_i(t) = \{ (v, \kappa) | A(t, v, \kappa) \cap B_i(t) \neq \emptyset \}. \quad (23)$$

$FVOCL_i(t)$ and $FVOCL_i$ can be determined as well, according to (21) and (22).

The main difference is the following: the grazing points in $VOCL_i(t)$ ((v_{g1}, κ_{g1}) and (v_{g2}, κ_{g2})) do not represent grazing cases any more if the obstacle is moving since the position of the grazing points depends on the motion of the obstacle.



To analyze VOCL and the grazing points, the following notations are used (see Figure 6):

- A time instant t_A is considered.
- The position of the moving obstacle at t_A is denoted by $(x_o(t_A), y_o(t_A))$. The angle between the positive x-axis and

the velocity vector $\mathbf{v}_o(t_A)$ of the obstacle is denoted by $\theta_o(t_A)$, the absolute value of the velocity is $v_o(t_A)$. $\omega_o(t_A)$ denotes the turning rate of the obstacle. The vector $\boldsymbol{\omega}_o(t_A)$ is parallel to the positive z-axis, if the direction of the rotation is positive, and to the negative z-axis in case of clockwise rotation.

- An input pair (v_A, κ_A) is selected.
- If a robot is moving with a velocity v_A on a circle with curvature κ_A for time t_A, its position $(x_r(t_A), y_r(t_A))$ and orientation $\theta_r(t_A)$ can be determined according to (8). The absolute value of the robot's velocity v_A remains unchanged during the circular motion, but the vector of the velocity $\mathbf{v}_A(t_A)$ changes, since the orientation of the robot is also modified. The turning rate of the robot is $\omega_A = \kappa_A\, v_A$. The direction of vector $\boldsymbol{\omega}_A$ is parallel to the positive z-axis, if $\omega_A > 0$, and to the negative z-axis if $\omega_A < 0$. The vector $\boldsymbol{\omega}_A$ is constant during the circular motion, since its direction and length (i.e. ω_A) are unchanged.
- The relative velocity of the robot and the obstacle at time moment t_A is denoted by

$$\mathbf{v}_{r|o}\,(t_A) = \mathbf{v}_A(t_A) - \mathbf{v}_o(t_A). \tag{24}$$

- The angle between the velocity vector $\mathbf{v}_A(t_A)$ of the robot and the relative velocity $\mathbf{v}_{r|o}(t_A)$ is:

$$\beta(t_A) = \operatorname{atan}\frac{v_o(t_A)\sin(\theta_r(t_A)-\theta_o(t_A))}{v_A - v_o(t_A)\cos(\theta_r(t_A)-\theta_o(t_A))}. \tag{25}$$

- $\mathbf{p}_{or}$ is a vector connecting the center of the obstacle to the center of the robot.
- The angle between the positive x-axis and $\mathbf{p}_{or}$ is:

$$\gamma(t_A) = \operatorname{atan}\frac{y_r(t_A)-y_o(t_A)}{x_r(t_A)-x_o(t_A)}. \tag{26}$$

The following two propositions follow directly from the definition of $VOCL_i(t)$.

Proposition 1:

$$\sqrt{\left(x_r(t_A)-x_o(t_A)\right)^2 + (y_r(t_A)-y_o(t_A))^2} \le r_o' \Leftrightarrow$$

$$(v_A, \kappa_A) \in VOCL_i(t_A). \tag{27}$$

If $|\kappa_A| \le \kappa_{max}$ and $|v_A| \le v_{max}$, then $(v_A, \kappa_A) \in FVOCL_i(t_A)$ is also true. Similar to [14], the boundary points of the set $VOCL(t)$ are denoted by $\delta VOCL(t)$.

Proposition 2:

$$\sqrt{\left(x_r(t_A)-x_o(t_A)\right)^2 + (y_r(t_A)-y_o(t_A))^2} = r_o' \Leftrightarrow$$

$$(v_A, \kappa_A) \in \delta VOCL_i(t_A). \tag{28}$$

For the grazing points (v_{g1}, κ_{g1}) and (v_{g2}, κ_{g2}) in case of a static obstacle still holds true that $(v_{g1}, \kappa_{g1}) \in \delta VOCL_i(t_A)$ and $(v_{g2}, \kappa_{g2}) \in \delta VOCL_i(t_A)$ where δ denotes the region boundary. But these points are not grazing points any more, if the obstacle is moving.

Theorem 3: (v_A, κ_A) is a grazing point in $VOCL_i(t_A)$ if

$$\theta_r(t_A) + \beta(t_A) - \gamma(t_A) = \frac{(2a-1)\pi}{2}, a \in \mathbb{Z}. \tag{29}$$

Proof: $\theta_r(t_A) + \beta(t_A)$ denotes the angle between the relative velocity $\mathbf{v}_{r|o}(t_A)$ and the positive x-axis. $\gamma(t_A)$ is the angle between the positive x-axis and $\mathbf{p}_{or}$ (the vector connecting the center of the robot and the obstacle). The condition for the grazing situation is that the relative velocity $\mathbf{v}_{r|o}(t_A)$ is perpendicular to $\mathbf{p}_{or}$ which is equivalent to the case where the relative velocity $\mathbf{v}_{r|o}(t_A)$ is tangent to the circle of the enlarged obstacle with radius r_o'.

Similar to [14], it can be proven that if $\mathbf{v}_{r|o}(t_A)$ is perpendicular to $\mathbf{p}_{or}$, the robot grazes the obstacle.

A point P is selected on the boundary of the robot. $\mathbf{p}_P$ is a vector which goes from the center of the robot to P. The velocity of the point P is:

$$\mathbf{v}_P(t_A) = \mathbf{v}_A(t_A) + \boldsymbol{\omega}_A \times \mathbf{p}_P. \tag{30}$$

Similarly, a point Q can be selected on the boundary of the obstacle. $\mathbf{p}_Q$ is a vector which goes from the center of the obstacle to Q. The velocity of Q reads:

$$\mathbf{v}_Q(t_A) = \mathbf{v}_o(t_A) + \boldsymbol{\omega}_o(t_A) \times \mathbf{p}_Q. \tag{31}$$

If the point P of the robot grazes point Q on the obstacle, vector $\mathbf{p}_{or}$ has to be parallel to $\mathbf{p}_Q$ and $-\mathbf{p}_P$. Moreover, grazing case can only occur, if the relative velocity of point P according to point Q is tangent to the circle of the obstacle, i.e. if it is perpendicular to $\mathbf{p}_{or}$ (and accordingly to $\mathbf{p}_Q$ and $\mathbf{p}_P$ as well). The relative velocity of P according to Q is:

$$\mathbf{v}_{P|Q}(t_A) = \mathbf{v}_P(t_A) - \mathbf{v}_Q(t_A) =$$

$$= \mathbf{v}_A(t_A) + \boldsymbol{\omega}_A \times \mathbf{p}_P - \mathbf{v}_o(t_A) - \boldsymbol{\omega}_o(t_A) \times \mathbf{p}_Q. \tag{32}$$

$\mathbf{v}_{P|Q}(t_A)$ is perpendicular to $\mathbf{p}_{or}$, if:

$$\mathbf{v}_{P|Q}(t_A) \cdot \mathbf{p}_{or} = 0$$

$$(\mathbf{v}_A + \boldsymbol{\omega}_A \times \mathbf{p}_P) \cdot \mathbf{p}_{or} - \left(\mathbf{v}_o + \boldsymbol{\omega}_o \times \mathbf{p}_Q\right) \cdot \mathbf{p}_{or} = 0. \tag{33}$$

Both $(\boldsymbol{\omega}_A \times \mathbf{p}_P) \cdot \mathbf{p}_{or}$ and $\left(\boldsymbol{\omega}_o(t_A) \times \mathbf{p}_Q\right) \cdot \mathbf{p}_{or}$ equal 0, hence one gets

$$\left(\mathbf{v}_A(t_A) - \mathbf{v}_o(t_A)\right) \cdot \mathbf{p}_{or} = 0, \tag{34}$$

which is equivalent to that $\mathbf{v}_{r|o}(t_A)$ is perpendicular to $\mathbf{p}_{or}$.■

Notice, that the obstacles can move on arbitrary paths with arbitrary velocity profiles. The algorithm only supposes that for every time moment $t_A \in [t_0, t_h]$ the positions and the velocity vectors of each obstacles are known.

5.3. Trajectories Avoiding Obstacles

If a goal position is given in the workspace of the robot, a motion should be planned which arrives to this point. During the motion, no collision should occur.

Fiorini and Shiller presented some heuristic search methods using Velocity Obstacles [6]. Applying these, one can easily select at each time moment a feasible velocity vector which moves the robot to the direction of the goal such that it avoids collisions with the obstacles. Using these heuristics alone does not guarantee to find an optimal solution. They designed an off-line global search method as well. A tree was generated for avoidance maneuvers using VO.

Shiller et al. presented avoidance maneuvers using NLVO [14]. These avoidance maneuvers can be used in a local or global motion planner as well.

The VOCL method can also be used in all above methods to plan the motion for car-like mobile robots to a given goal avoiding static and moving obstacles.

5.4. The Safest Solution

Another possibility for the selection of (v, κ) pair is to determine the safest motion. This method is called as the Safety Velocity Obstacles method (SVO).

The most important goal of the SVO method is to ensure the safest path for the robot during the motion.

This method checks, how far a possible $(v, \kappa) \notin FVOCL$ pair is from the nearest $FVOCL_i$ (name this distance VO_{dist}). If the distance is bigger than a predefined threshold d_{max}, then the value of VO_{dist} is set to this maximum distance.

Hence, a normalized distance can be defined:

$$VO_{norm} = \frac{VO_{dist}}{d_{max}} \tag{35}$$

where $VO_{norm} \in [0,1]$. After that the value of the cost can be defined as

$$VO_{cost} = 1 - VO_{norm}. \tag{36}$$

If the selected (v, κ) pair is far from the nearest $FVOCL_i$, then the value of VO_{norm} will be a big number (near to 1). So, such (v, κ) pair has to be chosen for the robot, where the value of VO_{cost} is minimal.

So, with the SVO method one can plan the safest motion in dynamic environment.

6. Simulation Results

An example for VOCL is presented here. A circular robot with radius $r_o = 0.9\text{m}$ is given. Its minimal turning radius is $R_{min} = 3\text{m}$ (i.e. $\kappa_{max} = \frac{1}{3}\text{m}^{-1}$) and its maximal velocity is $v_{max} = 7\frac{\text{m}}{s}$. Seven circular obstacles with different radii are presented in the workspace (see Figure 7). Four obstacles move (B_1, B_2, B_3, B_4) and there are three static obstacles (B_5, B_6, B_7). The velocity vectors and the paths of the dynamic obstacles are also depicted in Figure 7. The VOCL and FVOCL for $t_h = 10\text{s}$ is given in Figure 8. White areas represent (v, κ) pairs which correspond to collision-free motion for t_h.

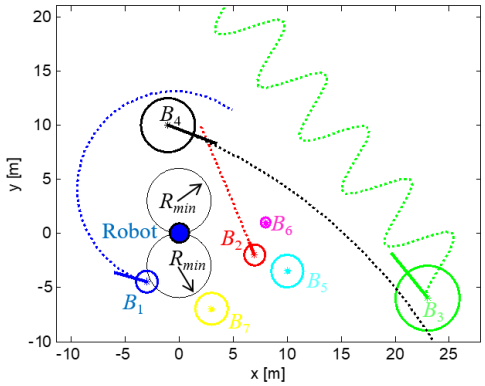
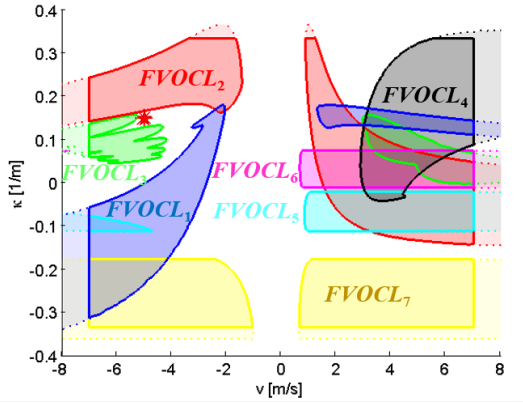


Figure 7. An example with seven obstacles at $t_0 = 0\text{s}$. The velocity vectors of dynamic obstacles are depicted by solid lines, the paths of the obstacles are denoted by dotted lines (with $t_h = 10s$).

A collision-free example is also presented here. $\kappa = 0.15\text{m}^{-1}$ ($R = 6.\dot{6}\text{m}$) and $v = -5\frac{\text{m}}{s}$ was selected. The corresponding (v, κ) point is depicted by a red star in Figure 8. The motion of the robot and the movement of the obstacles are depicted in Figure 9.



7. Conclusion

Velocity obstacles (VO) and non-linear velocity obstacles (NLVO) methods can be used to plan a collision-free motion for a planar robot moving among static and moving obstacles. VO supposes that the obstacles move on straight-lines with constant velocities. If the path of the obstacle is not a straight-line, NLVO can be applied. These methods determine a velocity vector for the robot, which results in a collision-free motion in a time-interval. Applying this velocity, the robot will move on a straight-line according to the direction of the selected velocity vector. Both methods assume that the position and the motion of the obstacles are known.

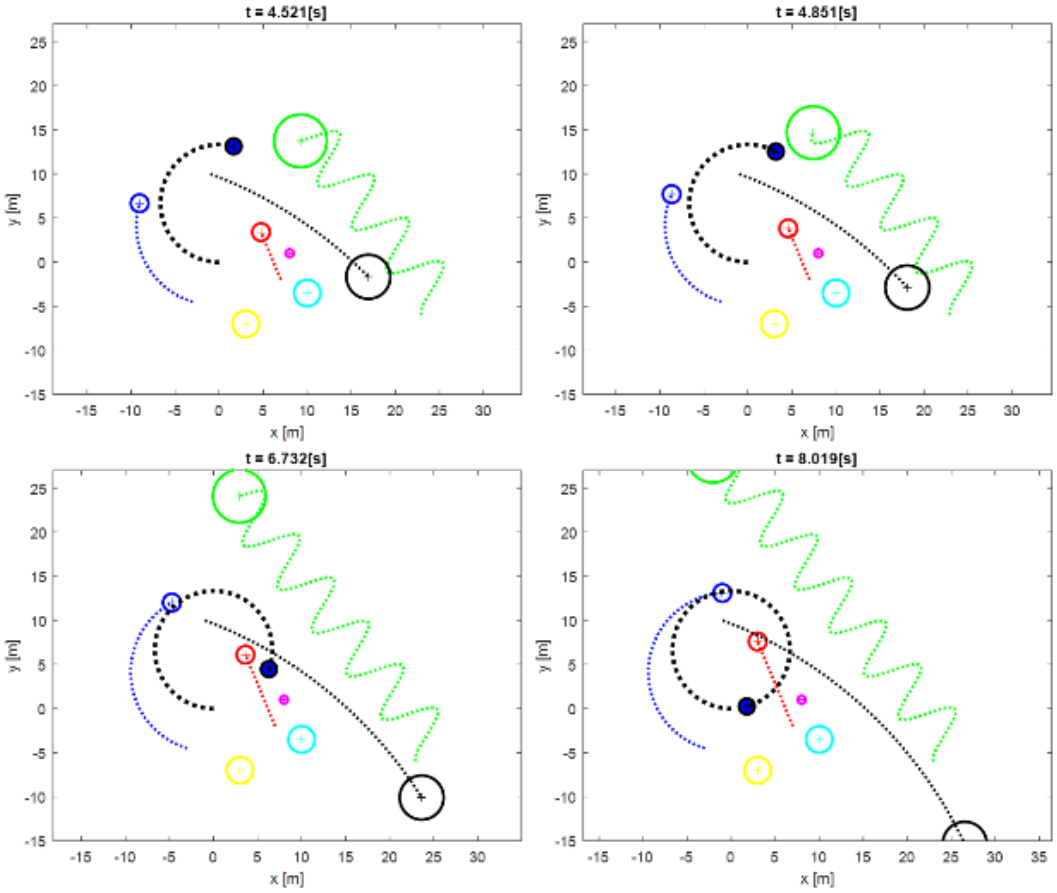


Figure 9. Motion of robot and obstacles in the example ($v = -5\frac{m}{s}$, $\kappa = 0.15m^{-1}$). The robot is depicted by a blue disc, and its path is shown by the thick black dotted line.

The method, presented in this paper is similar to NLVO but in this case the robot is a car-like mobile robot. This robot cannot move to arbitrary direction. The direction of its velocity vector is determined by its orientation. The robot can move on a straight-line according to its orientation or on a circular path. Hence, the presented VOCL method determines only the magnitude of the velocity vector and, additionally, the curvature of the circular path to follow to avoid collisions.

The next stage of this work is to implement the VOCL method for a car-like mobile robot, such that the collision-free motion is determined in real-time.

Our future goal is to take the non-circular shape of the robot (e.g. car-like rectangle) also into consideration during the construction of VOCL.

[22]. In this case a continuous curvature path could be determined. The drawback is that the computational complexity would be larger.

References

[1] E. Gincsainé Szádeczky-Kardoss, B. Kiss, "Velocity obstacles for Dubins-like mobile robots" in 25th Mediterranean Conference on Control and Automation (MED), Valletta, Malta, 2017. https://doi.org/10.1109/MED.2017.7984142

[2] J. C. Latombe, Robot motion planning, Kluwer, Boston, 1991.

[3] S. M. LaValle, Planning algorithms, Cambridge University Press, 2006.

[4] K. Kant, S. W. Zucker, "Toward efficient trajectory planning: The path-velocity decomposition" Int. J. Robot. Res., **5**(3), 72–89, 1986. https://doi.org/10.1177/027836498600500304

[5] E. Szádeczky-Kardoss, B. Kiss, "Motion planning in dynamic environments with the rapidly exploring random tree method" Int. Rev. of Automatic Control, **1**(1), 109–117, 2008.

[6] P. Fiorini, Z. Shiller, "Motion planning in dynamic environments using velocity obstacles" Int. J. of Robot. Res., **17**(7), 760–772, 1998. https://doi.org/10.1177/027836499801700706

[7] M. Seder, I. Petrovic, "Dynamic window based approach to mobile robot motion control in the presence of moving obstacles" in International Conference on Robotics and Automation, Roma, Italy, 1986–1991, 2007. https://doi.org/10.1109/ROBOT.2007.363613

[8] O. Khatib, "Real-Time obstacle avoidance for manipulators and mobile robots". Int. J. Robot. Res., **5**(1), 90–98, 1986. https://doi.org/10.1177/027836498600500106

[9] C. Qixin, H. Yanwen, Z. Jingliang, "An evolutionary artificial potential field algorithm for dynamic path planning of mobile robot" in IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 3331–3336, 2006. https://doi.org/10.1109/IROS.2006.282508

[10] P. Shi, J. N. Hua, "Mobile robot dynamic path planning based on artificial potential field approach" Adv. Mat. Res., **490–495**, 994–998, 2012. https://doi.org/10.4028/www.scientific.net/AMR.490-495.994

[11] D. Fox, W. Burgard, S. Thrun, "The dynamic window approach to collision avoidance" IEEE Robot. Autom. Mag., **4**(1), 23–33, 1997. https://doi.org/10.1109/100.580977

[12] T. Fraichard, H. Asama, "Inevitable collision states – a step towards safer robots?" Adv. Robotics, **18**(10), 1001–1024, 2004. https://doi.org/10.1163/1568553042674662

[13] L. Martinez-Gomez, T. Fraichard, "Collision avoidance in dynamic environments: an ICS-based solution and its comparative evaluation" in IEEE International Conference on Robotics and Automation, Kobe, Japan, 100–105, 2009. https://doi.org/10.1109/ROBOT.2009.5152536

[14] Z. Shiller, F. Large, S. Sekhavat, C. Laugier, "Motion planning in dynamic environments: Obstacles moving along arbitrary trajectories". in IEEE

International Conference on Robotics and Automation, Seoul, South Korea, 3716–3721, 2001. https://doi.org/10.1109/ROBOT.2001.933196

[15] C. Moon, W. Chung, "Trajectory time scaling of a mobile robot to avoid dynamic obstacles on the basis of the INLVO" Adv. Robotics, **27**(15), 1189–1198, 2013. https://doi.org/10.1080/01691864.2013.819604

[16] Y. Kuwata, M. T. Wolf, D. Zarzhitsky, T. L. Huntsberger, "Safe maritime autonomous navigation with COLREGS, using velocity obstacles" IEEE J. Oceanic Eng., **39**(1), 110–119, 2014. https://doi.org/10.1109/JOE.2013.2254214

[17] J. Snape, J. v. d. Berg, S. J. Guy, D. Manocha, "The hybrid reciprocal velocity obstacle". IEEE T. Robot., **27**(4), 696–706, 2011. https://doi.org/10.1109/TRO.2011.2120810

[18] D. Wilkie, J. v. d. Berg, D. Manocha, "Generalized velocity obstacles". in IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 5573–5578, 2009. https://doi.org/10.1109/IROS.2009.5354175

[19] W. A. Wolfe, "Analytical design of an Ackermann steering linkage" J. Eng. Ind.- T. ASME, 11, 11–14, 1959.

[20] L. E. Dubins, "On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents" Am. J. Math., **79**(3), 497–517, 1957. https://doi.org/10.2307/2372560

[21] J. A. Reeds, L. A. Shepp, "Optimal paths for a car that goes both forwards and backwards". Pac. J. Math., **145**(2), 367–393, 1990

[22] T. Fraichard, A. Scheuer, R. Desvigne, "From Reeds and Shepp's to continuous-curvature paths" in International Conference on Advanced Robotics, Tokyo, Japan, 585–590, 1999.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 234-238 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Medium Voltage Microgrid Test Setup and Procedures Implemented on a Real Pilot Project

Bruno Alberto Pacheco*,[1], Marcos Aurelio Izumida Martins[1], Cesare Quinteiro Pica[1], Nilo Rodrigues[2]

[1]*CERTI Foundation, Sustainable Energy Center, Florianópolis - SC, Brazil*

[2]*ENEL, Research & Development, Fortaleza - CE, Brazil*

ARTICLE INFO

Article history:
Received: 13 November, 2017
Accepted: 15 January, 2018
Online: 30 January, 2018

Keywords:
Microgrid
Energy Storage System
Distributed Generation

ABSTRACT

This paper presents various concepts related to the application of a microgrid pilot project in a residential condominium at Fortaleza / CE - Brazil, such as battery energy system, renewable and distributed generation, islanding recloser and all different units using interface based on power electronics. This papers main objective is to create information about microgrid operation and the interaction between its main equipment, such as power converters, utility energy distribution system and control units responsible for algorithms and changes in microgrids operation mode. This information is important for understanding the need for a test setup construction. To perform the test procedures, a temporary setup in a controlled environment within the microgrid is proposed. During the test periods, intentional power outages are required to evaluate the operating mode switching on each unit. The test setup described in this paper aims to mitigate the tests effects on other residential units inside the condominium.

1. Introduction

Typical electric power systems are known to operate by transmitting energy in bulk from large generating plants over long distances to large load centers. With the introduction of distributed generation and the growing evolution of power electronics for the renewable generation sources of electricity, such as solar and wind, this scenario of operation has been changing gradually, leaving behind the centralized form of generation and becoming now decentralized, producing energy closer to consumers [1, 2].

In this context, microgrids present a new concept of power generation and distribution. Microgrids are utility power grid subsystems composed of a set of distributed energy resources, controllable electric loads, energy storage elements and power control system. Microgrid main characteristic is the ability to operate in island conditions, i.e., operate isolated from the main electricity distribution grid preserving an energy quality for microgrid connected consumers [3-5].

In addition to both connected and islanded modes, this microgrid, under analysis, presents a third operation mode referred to as Maintenance Mode. This operation mode is activated when a fault, such as short circuit, happens inside the microgrid, and consequently making it impossible to operate islanded. Therefore, maintenance mode becomes essential while aiming towards a safe microgrid operation. Recent studies [6, 7] also described the importance of such mode and its benefits for microgrid operations.

While in maintenance mode, the microgrid cannot be powered by any of its distributed energy resources (DER), i.e., the central energy storage (grid-forming power converter) and its internal sources of distributed generation (grid-feeding power converter) are unable to generate energy. Maintenance mode is also triggered with a power distribution company request, followed by command from the operation center, due to maintenance in the local medium voltage distribution grid, as it is procedure to disable the microgrid and ensure a complete absence of power at the condominium internal distribution grid.

To perform real tests of the control system operation mode switching and interaction with the equipment, a test setup for the pilot microgrid is proposed, aiming to mitigate the impact on consumers loads. In this paper, the microgrid project in analysis will be presented first and then the proposed test setup and procedures.

2. Microgrid Architecture

This microgrid under study is located in a residential condominium at Fortaleza / CE – Brazil. The condominium is connected to the utility through medium voltage grid.

*Corresponding Author: Bruno Alberto Pacheco, CERTI Foundation, Email: bap@certi.org.br

The Microgrid is divided into several units according to their characteristics and devices. This unit delimitation allows a better understanding of microgrid operation. Those units are shown in Figure 1.

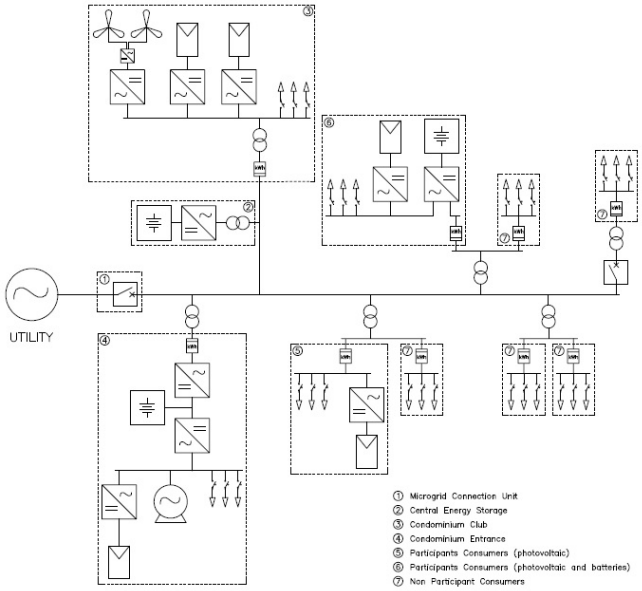


Fig. 1. Architecture of the real microgrid pilot project.

The main microgrid units are presented as it follows:

2.1. Central Energy Storage

This unit is composed by an energy storage system along with its power conversion and control devices. With a storage capacity of 105kWh and 250kW of power, this system is able to supply power for the microgrid while operating in islanded mode [8]. Figure 2 displays a simple electric diagram for this unit.

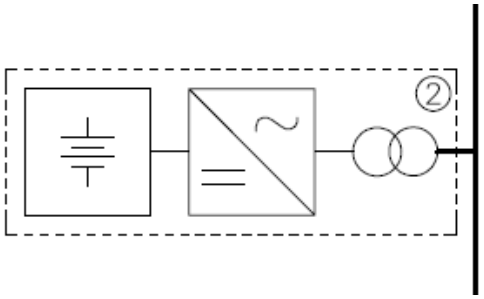


Fig. 2. Central Energy Storage.

The central storage system features a lithium-ion battery bank, grid-forming power converter, 380/13800 V power transformer, switchgear and their respective control and protection equipment.

During islanded mode, this unit function is to provide power for the loads and a voltage reference for all renewable distributed generation. While operation at connected mode, this unit is recharged and can also be discharged providing grid support services.

2.2. Consumers Units

Microgrids consumers are the systems main loads. They are strategically chosen residential consumers with distributed generation resources and some with local storage systems.

They are subdivided into Participant Consumers and Non Participant Consumers. Whilst operating in connected mode, both participant and non-participant act as loads for the system. During islanded operation, non-participant consumers are disconnected through smart meters. Therefore, only participant consumers remain connected at microgrid [8].

Figures 3 and 4 displays a simplified electrical diagram of those participant consumers. They are equipped with demand-response devices, solar generation, power inverters, smart meters, and some units are also equipped with a local storage system.

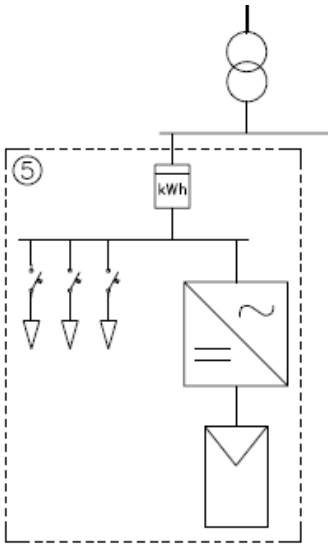


Fig. 3. Participant Consumers Units.

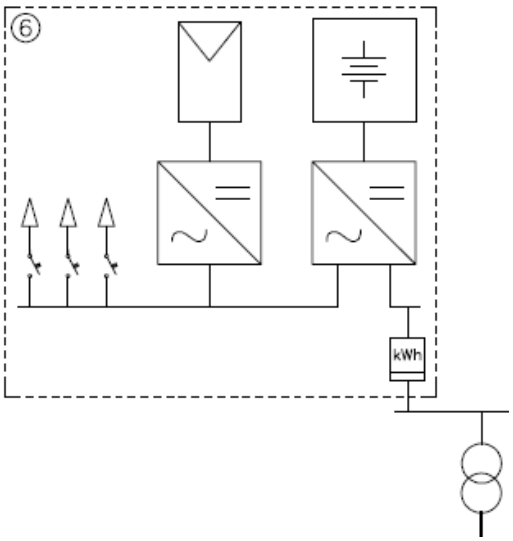


Fig. 4. Participant Consumers Units with a local storage system.

2.3. Club Unit

This residential condominium contains a club with common areas for its residents. The Club Unit (Figure 5) is also a microgrid consumer, however it has a larger distributed generations, composed with 24kWp of photovoltaic panels, as well as 7kW of wind generation.

Like the participating consumers, the Club is endowed with controllable loads through the demand-response devices. Another

important detail is that this unit is connected at microgrid in medium voltage and has its own substation.

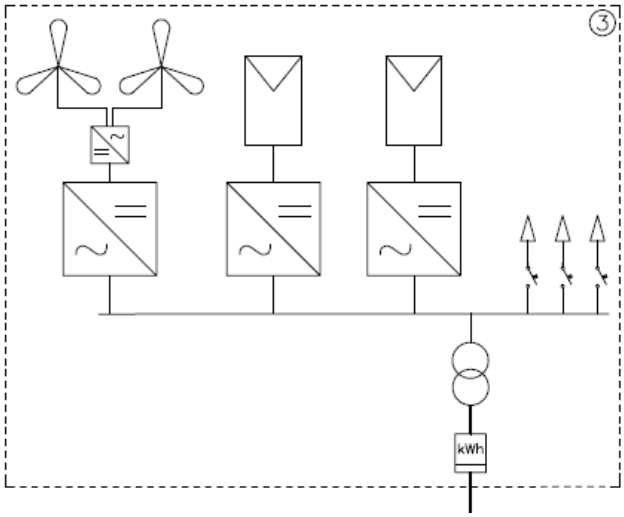


Fig. 5. Club Unit.

2.4. Condominium Entrance

This unit contains critical loads for condominium security and safety functions. Due to this fact, it is equipped with a diesel generator as a back-up unit.

A back-to-back converter is employed as a connection interface with the utility grid, allowing for smooth operation during power outage scenarios. Furthermore, this diesel generator can be used in a microgrid support scheme, relieving the system during a utility outage [8].

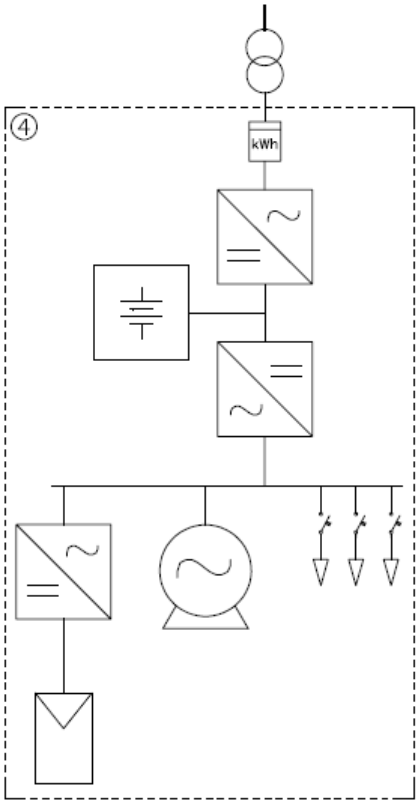


Fig. 6. Condominium Entrance.

This unit also has a solar generation system and controllable loads with demand-response devices as shown in Figure 6.

2.5. Microgrid Connection Unit

The Microgrid Connection Unit (Figure 7) consists of a medium voltage recloser and a set of control equipment for an energy outage detection in the distribution system, coordination of islanding and microgrid reconnection to the distribution system.

This unit is responsible for controlling the microgrid operation mode changes by informing other control units. It is also the communication interface with the utility grid SCADA (Supervisory Control and Data Acquisition) system.

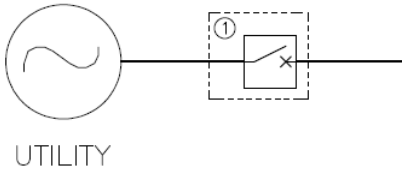


Fig. 7. Microgrid Connection Unit.

3. Microgrid Test Setup

In order to validate microgrid safe operation during its transitions through connected, islanded and maintenance modes, a power setup is proposed for conducting systemic tests (Figure 8).

The premise of this setup is to use the Club Unit area for real islanding, synchronization and reconnection microgrid tests. Since the Club Unit has a medium voltage connection, which is independent from the other condominium derivation line, any power supply interruption inside this area will not affect other microgrid units, reducing the inconvenience for all condominium residents.

As mentioned before, the Club unit has a larger amount of distributed generation and controllable loads. That can be used as a controlled test environment by simulating the residents homes energy consumption and generation. Thus, real tests of operation mode switching can be conducted in a safe way using the Club unit infrastructure.

In connected mode, the utility grid supplies the power for the loads and sets a voltage reference for the power electronics equipment of the generating units. The Central Energy Storage behaves like a grid-feeding power converter in this mode, only providing utility grid support when requested.

3.1. Islanding Tests

For islanding tests, the Microgrid Connection Unit (① in Figure 8) detects an outage in the distribution system due to under voltage protections. It then proceeds to disconnect the microgrid from the utility and sends a command to the Central Energy Storage, requesting a change in its mode of operation. The use of the temporary recloser (② in Figure 8) installed in the test setup input derivation is essential, in order to simulate an outage occurring when an intentional opening of this equipment is performed.

After disconnecting and receiving the command from the Microgrid Connection Unit, the Central Energy Storage operates as a grid-forming power converter, assuming the role of the main power supply of the microgrid and providing voltage reference for the generation interfaces.

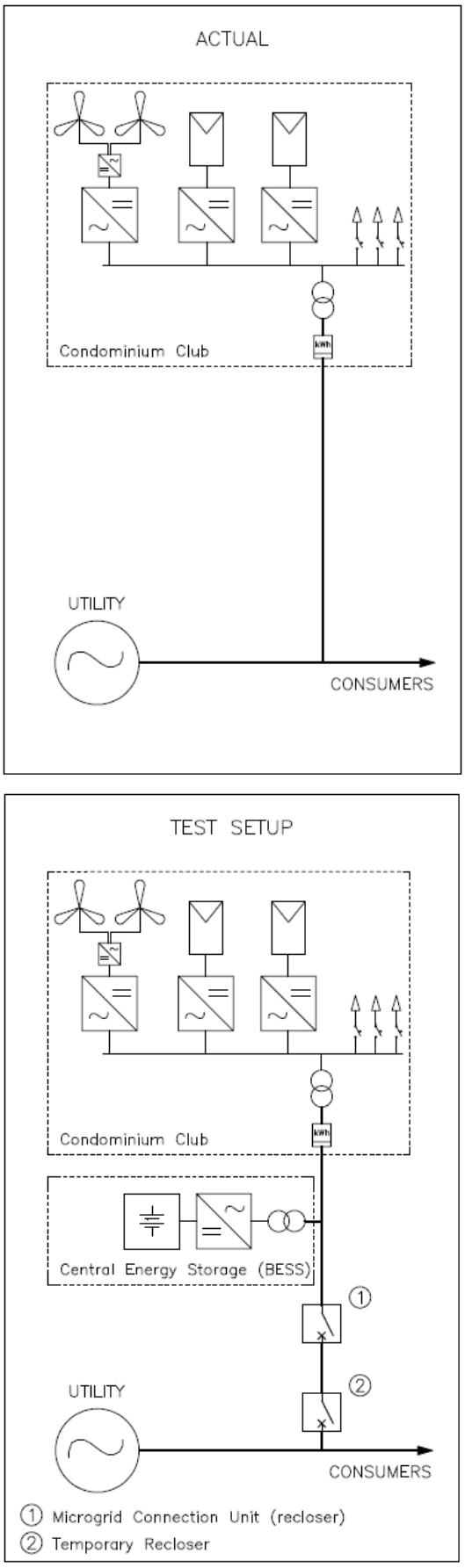


Fig. 8. Actual setup and Microgrid test setup.

3.2. Reconnecting Tests

When reconnecting the microgrid to the utility grid, the temporary recloser (② in Figure 8) is closed simulating the utility grid stable return after an outage.

The Microgrid Connection Unit detects the energy utility supply return while the Central Energy Storage synchronizes with the utility voltage reference. Microgrid is only reconnected to the main electricity distribution grid after synchronization is achieved.

Before the microgrid is connected to the utility grid, some of the indicators and parameters need to be checked, and the best time to close static switch should also be judged. This process is very important that can prevent the dangerous of transient phenomena during connecting the microgrid to utility grid. The good time to close the switch is when the voltage across the switch has to be very small (ideally zero), and the grid's frequency is higher than the microgrid's frequency. Because the current flowing through the switch is minimum, and the power flow direction is from utility grid to the microgrid in such a situation.

A safe and reliable grid-connection process will require to meet the following three conditions: a) the voltage through the static switch must be very small. b) microgrid operating frequency must be slightly smaller than the grid frequency. c) utility grid voltage must be ahead of the microgrid voltage [9].

3.3. Maintenance Tests

Maintenance mode is also tested in this setup with a request through utility operation center command. In such case, it must be verified that the Central Energy Storage is disconnected from the grid, not powering the microgrid when necessary.

Another advantage for this test setup is that all equipment employed will be used in the final installation. In this procedure it is possible to test the interaction between connected equipment in medium voltage, control system and power electronics interfaces. In addition, the Central Energy Storage is tested at its final installation place allows for other control and protection details to be verified.

4. Conclusions

In this work, the microgrid architecture on a pilot project at a residential condominium is presented, along with its main units, equipment and components according to its functionalities and technical characteristics.

Through the proposed test setup it is possible to obtain performance indicators on microgrid operating mode transitions and grid-forming power converter, as well as the integration between the control and protection equipment.

With the test setup execution, all operation modes shall be validated, along with transitions among them. The test results will improve the teams acknowledge around a correct microgrid functioning and a validation of its infrastructure and equipment.

Acknowledgment

The authors thank the Research Program and Technological Development of the electricity sector regulated by ANEEL and Enel Distribuição Ceará for the financial support to the project. This work is related to the project "Development Application of Pilot Microgrid Power Distribution with Distributed Generation

and Commercial Operations Model" under PD- 0039-0073 / 2014 ANEEL.

References

[1] H.M.A. Antunes, S.M. Silva and B.J.C. Filho, “Análise e Operação de uma Microrrede de Energia Elétrica,” VI Simpósio Brasileiro de Sistemas Elétricos, 2016.

[2] Gang HU et al. “Study on Modeling and Simulation of Photovoltaic Energy Storage Microgrid” 4th International Conference on Information Science and Control Engineering, p.692-695, 2017.

[3] R. H. Lasseter, “MicroGrids,” IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No.02CH37309), vol. 1, no. C, pp. 305–308, 2002.

[4] Daniel E. Olivares “A Centralized Energy Management System for Isolated Microgrids.” IEEE Transactions os Smart Grid, VOL. 5, July 2014.

[5] Benedetto Aluisio et al. “Embedding Energy Storage for Multi-Energy Microgrid Optimal Operation”, 2017.

[6] Sundari Ramabhotla et al. “Operation and Maintenance Cost Optimization in the Grid Connected Mode of Microgrid.” 2016 Ieee Green Technologies Conference, Kansas City, p.56-61, 2016.

[7] Richard B. Jones “How Reliable is your Microgrid: An insurer’s perspective on risk drivers for ditributed resources.”, Reston, July 2015.

[8] M.A.I. Martins et al., “Design and Implementation of a Microgrid Power Management Unit Using a Back-To-Back Converter in a Residential Condominium Connected at Medium Voltage,” COBEP SPEC IEEE Power Electronics Society, 2015.

[9] Meiqin Mao, Yinzheng Tao, Liuchen Chang, Yongchao Zhao and Peng Jin, "An intelligent static switch based on embedded system and its control method for a microgrid," IEEE PES Innovative Smart Grid Technologies, Tianjin, 2012, pp. 1-6.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 239-243 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Approximate method of analysis of log-periodic antennas with in-phase currents

Boris Levin*

Holon Institute of Technology, Israel

ARTICLE INFO

Article history:
Received: 23 August, 2017
Accepted: 07 January, 2018
Online: 30 January, 2018

Keywords:
Antenna theory,
Directional antennas,
In-phase current distribution,
Method of electrostatic analogy,
Numerical calculation

ABSTRACT

Log-periodic antennas have high electrical characteristics over a wide frequency range. The large length of the antennas is their significant disadvantage caused by the fact that each active region consisting of several radiators operates in a narrow frequency range. To extend this range, it is proposed to use radiators with concentrated capacitive loads, providing in-phase currents with a linear amplitude distribution law. An approximate calculation of the fields and directional characteristics of a complex multielement radiating structure is performed using a new method of electrostatic analogy based on the assumption that the ratio of emf in the radiator centers is equal to the charge ratio on the conductors of the electrically neutral system. The problem of optimization of considered antennas by methods of mathematical programming is formulated.

1. Introduction

The use of concentrated loads included along the axis of a linear antenna allows solving the synthesis problem, i.e., to obtain an antenna with required characteristics, more precisely with characteristics closest to required [1]. The special case of this problem - the creation of an antenna that provides a high level of matching in a wide frequency range and a maximum of radiation in a plane perpendicular to the axis of the radiator – has a great practical importance.

A conventional linear radiator does not satisfy these requirements. A reactive component of its input impedance is great everywhere except of a vicinity of a series resonance. When the length of a radiator arm is greater than 0.7λ (λ is a wavelength), the radiation in the plane perpendicular to its axis decreases, since anti-phase segments form on the current curve.

The task of creating a wide-range radiator is to determine the types and quantities of loads that provide high characteristics of the antenna in the operating frequency range. This task was solved by applying concentrated capacitive loads (capacitors) installed along the axis of the radiator and forming an in-phase current with a given law of the amplitude distribution along the antenna. The solution was based on an understanding of the advantages of in-phase current distribution and Hallen's hypothesis on the use of capacitive loads, the values of which vary along the radiator axis in accordance with a linear or exponential law. The chosen approach confirmed the Hallen's hypothesis and demonstrated the effectiveness of the proposed approximate methods of calculating capacitive loads: the method of impedance long-line and the method of a long-line with loads. The loads calculated by these methods were used as the initial values for numerical solution of the problem by methods of mathematical programming.

Since the placement of concentrated capacitive loads along the radiator allows to obtain not only a high level of matching of the antenna with the cable, but also high directional characteristics in a wide frequency range, the application of these loads had been considered in radiators – first in a linear antenna and then in directional one.

It should be emphasized that the optimization problem, as a rule, is solved in two stages. At the first stage an approximate method of analysis, based on physical representations about the nature of the problem, is developing. At the second stage, approximate results are used as initial values for the numerical solution of the problem by methods of mathematical programming.

The proposed method of an approximate analysis of antennas characteristics allowed us to calculate the directional patterns of radiators with a given current distribution and to compare performances of antennas, in which various laws of current distribution are realized. This is, firstly, the linear metal antenna with a sinusoidal current distribution along its axis, and secondly, the linear antenna with an in-phase current and with a linear or exponential law of an amplitude distribution. Calculations show that the use of linear antennas with capacitive loads makes it possible to obtain in the wide frequency band the desired shape of the radiation pattern, which provides a substantial increasing the communication distance.

*Boris Levin, Email: levinpaker@gmail.com

To evaluate this shape, a special parameter has been introduced - the pattern factor, equal to an average radiation level in a given range of vertical angles, for example, at angles θ from 60° to 90°. This parameter shows in fact what part of the signal is propagating along the earth's surface, i.e., it is useful for increasing the communication distance. With the help of the proposed method, it is shown that the in-phase current distribution in linear and V-radiators provides not only a high level of matching, but also the high directivity and smooth variation of this directivity over a wide frequency range [2].

For an approximate calculation of the fields of complex multielement radiating structures, the method of electrostatic analogy was proposed. It is based on the assumption that the ratio of emf in the radiators centers is equal to the ratio of the charges located on the conductors of electrically neutral system. This allows proposing a simple and effective procedure of calculating directional characteristics of director and log-periodic antennas with concentrated capacitive loads. The calculation results confirm the possibility of increasing the frequency range of the director antennas and decreasing the length of the log-periodic antennas.

The results of applying this method to calculating the directional characteristics of director antennas are given in [3]. They are compared with the characteristics of director antenna, described in [4] and consisting from four metal elements: active radiator, reflector and two directors. When optimizing the metal antenna, the mathematical programming method was used. The application of the approximate method of electrostatic analogy to a metal antenna confirmed the results obtained in [4]. With the help of the presented procedure, it was shown also that in-phase current distribution in radiators of the director antennas provides the higher directivity in a wide frequency range.

The remainder of this paper is organized as follows. In Section II, the proposed method of analysis is applied to the calculation of fields and directional characteristics of an active region of the log-periodic antenna. The numerical results obtained in Section III for such antennas with in-phase currents show that the number of such regions can be substantially reduced. The section IV contains conclusions, which emphasize a theoretical interest and practical benefits of the work.

2. Method of Analysis

In accordance with a well-known method of analysis of a log-periodic antenna [5], we consider an active region of this antenna, consisting of three radiators (Fig. 1), and find fields of these radiators when emf *e* is located in the center of the middle radiator. In accordance with Kirchhoff' law other generators must be short-circuited. As shown in [5], this means that emf's in the centers of other radiators and at the ends of sections of a two-wires distributive long line in the points of these radiators location are shorted also. We regard that the arm length of the middle radiator is $L_0 = \lambda/4$, the arm length of the left (longer) radiator is equal to $L_1 = \lambda/(4\tau)$, and the arm length of the right (shorter) radiator is $L_2 = \lambda\tau/4$, where τ is a denominator of a geometric progression, according to which the radiators' dimensions are changed. A magnitude σ is the other antenna parameter, equal to $\sigma = 0.25(1-\tau)\cot\alpha$, where α is an angle between an antenna axis and a line passing through radiators ends. The value σ is the number of wavelengths between the half-wave radiator and the smaller neighboring radiator: $b_2 = \sigma\lambda$.

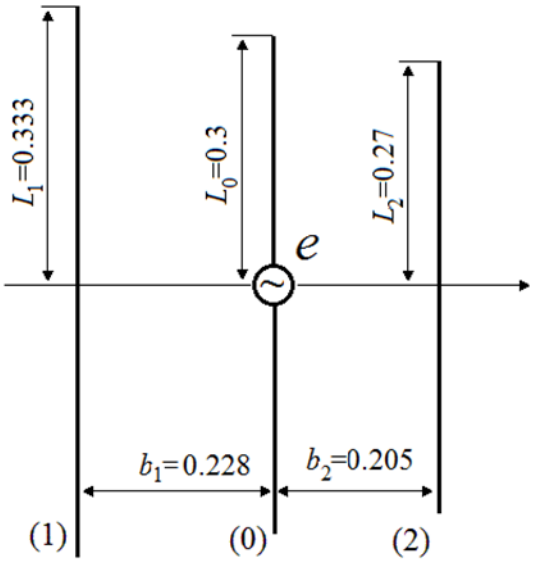


Figure 1: Antenna of three radiators

Respectively b_1 is equal to $\sigma\lambda/\tau$. As stated in [6], generalization of information in the literature leads to the conclusion that the minimum changes in the electrical characteristics of the log-periodic antenna with metal dipoles within the frequency range from f to $f\tau$ occur when $\sigma/\tau = 0.19$. We regard that this relation is valid for the considered antenna and that value τ for certainty is equal to 0.9. Then $L_1 = L_0/0.9 = 0.278\lambda$, $L_2 = L_0 \cdot 0.9 = 0.225\lambda$, $\alpha = 0.146$, $b_1 = 0.19\lambda$, $b_2 = 0.171\lambda$.

As already mentioned, we use the method of electrostatic analogy of two structures (of currents and charges), located on the same conductors, as an approximate method. We are assuming that the ratio of emf's in the radiator centers is equal to the ratio of charges located on these conductors. The positive charge, equal to Q_0, is placed on the conductor *0* (active radiator). The charges of other conductors *i* are negative. They are equal to $-Q_i$, and their sum is equal to $\sum_{i=1}^{2}(-Q_i) = -Q_0$, i.e., the sum of all charges is zero, and the conductors form an electrically neutral system. In this system

$$Q_i/Q_0 = C_{0i}\Big/\sum_{i=1}^{2} C_{0i}, \qquad (1)$$

where C_{0i} is the partial capacitance between conductors *0* and *i* (see, for example, [7]). From (1) it follows that the charges of the conductors *i* are directly proportional to the partial capacitances C_{0i} between these conductors and the conductor *0*. The mutual capacitance C_l per unit length of two parallel infinitely long conductors of the radius *a* is equal to

$$C_l = \pi\varepsilon/\ln(b/a), \qquad (2)$$

where ε b is the distance between the conductors. Multiplying this value into the conductor length,

we find in a first approximation the partial capacitance C_{0i} between these conductors.

We divide the active radiator (conductor *0*) into two parallel conductors (with numbers *01* and *02*), and the resulting structure of four conductors into two circuits: conductors with numbers *01* and *1* in the first scheme, and conductors with numbers *02* and *2* in the second scheme. In accordance with the method of electrostatic analogy of the two structures, i.e., in accordance with the physical content of the problem, we will assume that the ratio of the emf in the centers of the wires *01* and *02* is equal to the ratio of the partial capacitances C_{01} and C_{02}, i.e., $e_1 = 0.52e, e_2 = 0.48e$.

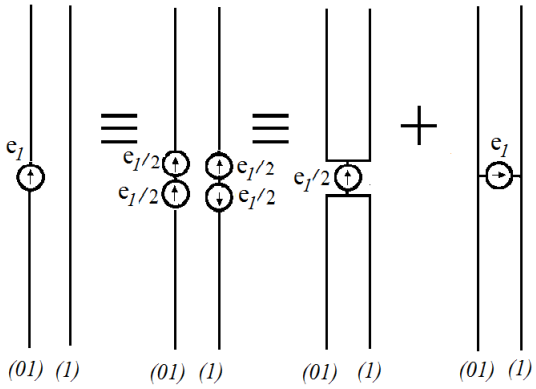


Figure 2: Successive transforming a structure of two parallel vertical conductors

Fig. 2 demonstrates a scheme for successive transforming a structure of two parallel vertical conductors (*01* and *1*), one of which is excited at the center, into a structure consisting of a two-conductor dipole and two open transmission lines. The scheme is based on the theory of a folded antenna (see for example, [1]). As can be seen from this figure, the current at the center of each dipole conductor and the current at the same point of each transmission line are respectively equal

$$J_{1d} = e_1/(4Z_{1d}), J_{1l} = e_1/(2Z_{1l}), \tag{3}$$

where Z_{1d} is the input impedance of this dipole, Z_{1l} is the input impedance of each long line. Accordingly, the current J_{01u} of the upper arm of the conductor *01* is equal to the difference of these currents, and the current of the lower arm of the conductor *01* - to the sum:

$$J_{01u} = J_{1d} - J_{1l}, J_{01l} = J_{1d} + J_{1l}. \tag{4}$$

The current J_{1u} of the upper arm of the conductor *1* on the contrary is equal to the sum of these currents, and the current J_{1l} of the lower arm of the conductor *1* – to their difference. The field amplitude and phase value of each radiator depend on its structure and location in an antenna. If the radiator arm is manufactured as the straight metallic wire, the current along it is distributed by the sinusoidal law

$$J(z) = J(0)\sin k(L - |z|). \tag{5}$$

Here z is a coordinate along the radiator axis, $J(0)$ is the current in its center, k is the propagation constant. In this case the far field of the radiator is equal to

$$E = \frac{AJ(0)}{Sin\theta}[\cos(kL\cos\theta) - \cos kL], \tag{6}$$

where $A = j\frac{60}{\sin kL}\cdot\frac{\exp(-jkR)}{R}$, R is the distance to the observation point. If concentrated capacitive loads located along the radiator axis allow realizing in it the in-phase current with an amplitude, distributed along this axis by the linear law,

$$J(z) = J(0)(L - |z|), \tag{7}$$

the far field of the radiator (see [2]) is equal to

$$E = \frac{AJ(0)\sin\theta}{\cos^2\theta}[1 - \cos(kL\cos\theta)]. \tag{8}$$

Consider an impact of the radiator location on the far field of the antenna by way of a specific example of the director antenna, shown in Fig. 1. The arm length of the middle (active) radiator is equal to 0.3 м, i.e., the wave length of the first (series) resonance is equal to 1.2 м. Radii of all conductors are the same and are equal to 0.001м. The magnitude τ is 0.9.

It is obvious that the maximal radiation of the antenna should be directed to the right, toward the radiator *2*. Since the radiator *1* is located from the left of the active radiator *0*, at a distance b_1 from it, its field lags behind the field of the active radiator, first on kb_1 in phase, that is by the time interval of the propagation of the signal from the active radiator to the passive radiator *1*, and, secondly, by $kb_1/\sin\theta$ in phase, i.e., by the time of the signal propagation in the opposite direction, from the radiator *1* to the active radiator (signal of the wire *1*, radiated at an angle θ, must come to the active radiator at the same angle θ, i.e., it must pass the distance $b_1/\sin\theta$, and not the distance b_1. The total change in phase is equal to $\psi_1 = -kb_1\cdot\frac{1+\sin\theta}{\sin\theta}$. Similarly, in the case of radiator *2*, this phase change is equal to $\psi_2 = -kb_2\cdot\frac{\sin\theta - 1}{\sin\theta}$.

The total field of the antenna structure shown in Fig. 1 with in-phase current distribution at angle θ on the base of aforesaid may be written in the form

$$E = \frac{AJ(0)\sin\theta}{2Z_{1d}\cos^2\theta}\sum_{i=1}^{2} e_i\{[1 - \cos(kL_0\cos\langle\theta - \theta_0\rangle)] + \exp(j\psi_i)[1 - \cos(kL_i\cos\theta)]\}. \tag{9}$$

The directivity magnitude is determined by the expression

$$D = |E(\pi/2)|^2 \Big/ \sum_{n=1}^{N}[|E(\theta_n)|^2\Delta\sin\theta_n], \tag{10}$$

where Δ is the interval between neighboring values θ_n, N is the number of these intervals between $\theta = 0$ and $\theta = \pi/2$.

3. Examples of Calcultations

The results of calculating the directivity magnitude for the structure, shown in Fig. 1, are given in Fig. 3 depending on an electrical length kL_0

$kL_0 = \pi$ to $kL_0 = 10\pi$.

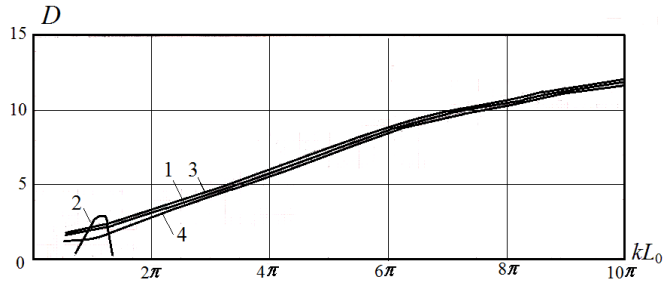


Figure 3: Directivity of structures of three radiators

Radiation of neighboring elements of log-periodic antenna with the arm length 0.27 m and 0.333 m respectively requires calculating fields in the structures presented in Fig. 4*a* and 4*b*. Results of these calculations are given in Fig. 3: curve 3 corresponds to Fig. 3*a*, curve 4 - to Fig. 3*b*. The directivity magnitudes in Fig. 3 for specific values kL correspond to the same frequencies, i.e., the same values of kL correspond to the elements of equal length in the all three schemes (for example, to the elements with the arm length 0.3 m). Fig. 3 shows that the directivity magnitudes at the same kL are close to each other. This is natural, since the directivity in each scheme increases slowly with increasing frequency. Small increasing the active radiator length causes at the same kL the small increase of the resonant wavelength, i.e., the small decrease of the resonance frequency and the directivity.

The obtained results show that each active radiator with in-phase current included in the structure of the log-periodic antenna provides high radiation directivity in a wide frequency range. Neighboring radiators have similar directivity magnitude. The direction of the radiation is the same. A signal propagates along the distributive line from short elements to the long dipoles (in Figures 1 and 4 to the left), the radiated signal propagates in the opposite direction. The total path difference is quite large. For example, for the signals radiated by a half-wave radiator and a smaller neighboring radiator it is equal to 0.38 λ, i.e., it is close to a half wavelength. Crossing wires of the distributive line in an interval between the elements permits to reduce dramatically this path difference.

Known log-periodic antenna with sinusoidal current distribution along the radiators have a property of automatic currents "cut-off", i.e., a separate antenna section (active region) radiates a signal in a narrow frequency band. Outside this band, outside the borders of the active region the signal decays rapidly. Wide frequency range is provided by a large antenna length, which is equal to the sum of the lengths of the active regions. Attempts to reduce the antenna length by disturbing a geometric progression and increasing a radiators number lead to a small dimension decrease and a sharp deterioration of electrical characteristics. The more effective methods are firstly a two-fold use of each active region by an application of linear-spiral radiators and secondly an employment of an asymmetrical log-periodic antenna with coaxial distribution line [6]. These methods allow decreasing the antenna length by 25-30%.

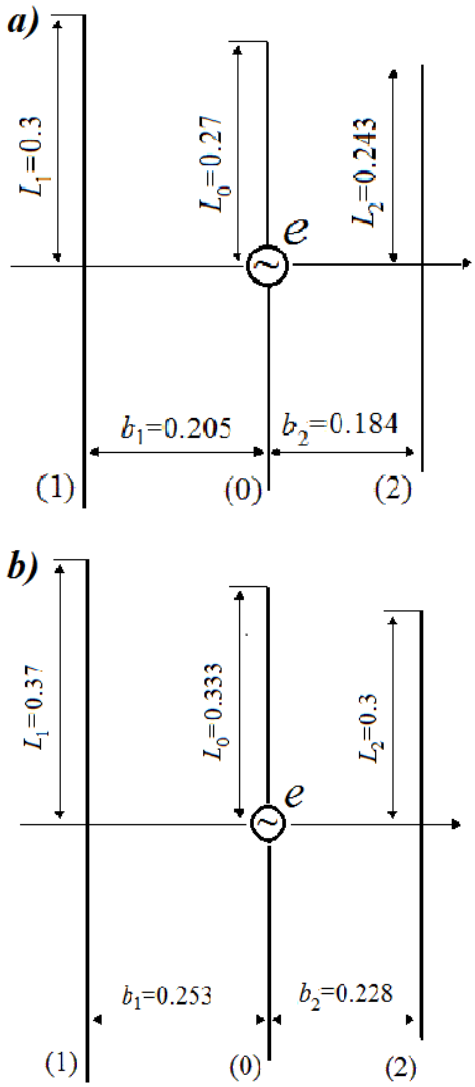


Figure 4: Structures of smaller (*a*) and greater (*b*) neighboring radiators

Replacement of the straight metal radiators by the radiators with concentrated capacitive loads allows obtaining a high directivity in a wide frequency range, using a simple structure of three radiators. Increasing the radiators number in the structure must allow to dramatically increasing its directivity.

4. Conclusion

In this work the following results are obtained:

1) The method of electrostatic analogy is proposed for calculating fields of complex multielement antennas. It is based on an assumption that a ratio of emf in radiators centers is equal to a ratio of charges located on conductors of an electrically neutral system. This method had been used for calculating directional characteristics of log-periodic antennas.

2) The given procedure permitted to compare the directional characteristics of log-periodic antennas with sinusoidal and in-phase current distribution in simple radiators. It was shown that replacement of the metal radiators by the radiators with concentrated capacitive loads allows providing, using a few

radiators, the high directivity in a wide frequency range, i.e., decreases sharply the antenna length. Increasing radiators number must allow increasing substantially the antenna directivity.

Results obtained by means of the method of electrostatic analogy may be used for solving optimization problem log-periodic antennas by methods of mathematical programming.

References

[1] 1. B. M. Levin. Antenna engineering. Theory and problems. A science publishers book: London, New York, 2017.

[2] 2. B. M. Levin. "Directional properties of linear and V-antennas" in 11 Intern. Confer. ICATT'17. Kiev (Ukraine), 2017, pp.104-109.

[3] 3. B. M. Levin. "Director antennas with in-phase currents" in 11 Intern. Confer. ICATT'17. Kiev (Ukraine), 2017, pp.110-113.

[4] 4. A. F. Chaplin, M. D. Buchazky, M. Yu. Mihailov. "Optimization of director-type antennas". Radiotechnics, 1983, no. 7, pp. 79-82 (in Russian).

[5] 5. R. L. Carrel. "The design of log-periodic dipole antennas. IRE Intern. Convention Record, part 1, 1961, pp. 61-75.

[6] 6. A. F. Yakovlev, A. E. Pyatnenkov. Wide-band directional antennas arrays from dipoles. S.-Petersburg: Research center of communication, 2007 (in Russian).

[7] 7. Yu. Ya. Iossel, E. S. Kochanov, M. G. Strunsky. Calculation of Electrical Capacitance. Leningrad: Energoisdat, 1981 (in Russian).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 244-251 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Measuring modifiability in model driven development using object oriented metrics

Nwe Nwe[*,1], Ei Thu[2]

[1]*Computer University, Monywa, Myanmar*

[2]*University of Computer Studies Mandalay, Myanmar*

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 02 January, 2018
Online: 30 January, 2018

Keywords :
Model transformation
Drools rule-based
Modifiability

ABSTRACT

Model driven development is an important role in software engineering. It consists of multiple transformation functions. This development is a paradigm for writing and implementing computer program quickly, effectively, at minimum cost and reducing development efforts because it transforms design model to object-oriented code. Our approach is rule-based model driven development in which textual Umple model is used as primary artifact and transformed to mobile applications. In this model driven development, evaluation of quality of transformation is critical. This paper has presented a set of metrics to assess the quality attribute of modifiability and evaluated using these object-oriented metrics. Results represent our approach achieves high efficiency in quality of modifiability.

1. Introduction

Changing with technology, mobile devices and mobile applications are necessary thing for every person and every sector. The burst on the availability of mobile devices is powering a growing mobile application. According to this fact, mobile applications, which are used in these devices, are critical in an industrial development. To fulfill the demand of these things, mobile application development preferably uses model-driven engineering than traditional software development process widely and effectively. It focuses on model for the development of software. There is lower the overall cost of building large internal applications, there is lower the risk of large application, speed time to build large applications and expand the pool of resources that can work on large application are main strategic objectives to use model driven development. Reduction of both direct and indirect development efforts, which enables scripters to contribute to enterprise development and enables task-oriented management of development are benefit of model driven development. It is a superset of model driven development because it goes beyond the traditional development.

In this case, the rule based model driven development of mobile application using Drool Knowledge-based Rule was presented in [1]. They also measured assessment of transformability using object oriented metrics. Drools Knowledge-based is a business rule management system with a forward and backward chaining inference based rules engines [2].

Moreover, this model-driven development is a development paradigm that uses model as the primary artifacts of the development process. It transforms source model to target model/code according to changing requirement and software reused more rapidly than traditional software development. A model transformation consists of multiple transformation functions. These transformation functions transform target language elements from the source language. Most of researchers concentrate on model to model transformation using intermediate meta-model or model to code transformation [3-6]. There is no quality, there is no efficiency in everything. Quality issue also change scale and become more important. The process transforms to new model or code related to quality of final software product and the quality of the model used to generate it. The consistency of source and target model and the assessment of quality of transformation is the critical issue in model transformation domains.

There are many attributes to evaluate quality in software engineering. Among them, most of these quality attributes can be applied to software artifacts in general. However, in [7] authors describe two quality criteria important in model transformation. They are transformability and modifiability. According to their

[*]Corresponding Author: Nwe Nwe, Email: nwenwemdy08@gmail.com

work, it is necessary to extend the assessment to other quality attribute of modifiability, maintainability and reusability using object-oriented metrics. Modifiability is the extent to which a model transformation can be adapted to provide different or additional functionality. The main reason for modifying a model transformation is changing requirements. Another reason is that the (domain specific) language in which the source and/ or target model are described which may be subject to changes. Modifiability captures the amount of effort needed to modify a model transformation. It is the combination of the modularity and reusability, an essential aspect of software engineering that promotes software maintainability. Moreover, it enables transform without affecting other parts of a program that are not directly connected to the changes. This paper is organized as follows: Section II explains the basic concepts of related work. Section III presents the contribution of rule based model transformation and quality attributes of modifiability. Section IV describes our experimental results and comparison results. Finally, section V concludes our approach and evaluation of modifiability using object oriented metrics.

2. Related work

There is increasing attention towards the generation of source code from modeling languages. Several researchers propose model driven approach for the different aspects of mobile applications. The model driven development of mobile applications using Drools knowledge-based Rule was describes in [1]. They developed mobile application by applying Drool rule based. Their work is closely related with JUSE4 android application [8,9]. Moreover, they attempted to address the consistency of source and target model and the assessment of transformability by measuring the accuracy of consistency between source and target model and assessing the transformability using object oriented metrics. According to their work, it is necessary to extend the assessment to other quality attributes of model driven development using object-oriented metrics. Authors performed a comparative study on C++, C# and Java programs using object-oriented metrics in [10]. It consists of class size, complexity, coupling cohersion, inheritance, encapsulation, polymorphism and reusability.

An evaluation of the quality of model transformation was defined in [11]. They made the quality of model transformation measurable. They presented the quality attributes and a set of metrics to assess these quality attributes. A calculation of metrics values using the same set of standard metrics for three software system of different sizes was described in [12].

In [5], authors presented quality goal in MDE and states that the quality of models is affected by the quality of modeling languages, tools, modeling processes, the knowledge and experience of modelers and the quality assurance techniques applied. In [3, 4], authors defined the meta-model and model transformation rule for model driven android application development. In [13], authors also defined ATLAS transformation rules for UML sequence diagram to generate enterprise java bean code (EJB). In [6], authors presented enhance code generation tool for android source code based on UML class and sequence diagram. In [14], authors specified meta-model with Ecore and transformation rules with Xpand templates for entity relationship diagram to generate android SQLite database model.

In model driven transformation, the approaches are quite different in their respective use of input model. Most of these approaches are based on graphical modeling or textual modeling languages. In contrast to our approach, the previous approaches applied pre-defined meta-modeling while our approach automatically parses and extracts syntax form input model of Umple [15]. Moreover, our approach has specified transformation rules in object pattern matching approach. JUSE4Android is also based on textual modeling languages. Unlike our approach, it is adding annotation into JUSE model and transform into android source according to the predefined meaning of annotation. Therefore, they generated source code contains some more files in their project. The authors [16, 17] have proposed the approach for empirical evaluation of model driven engineering in multiple dimensions. Their case studies include qualitative (expert judgements) and quantitative data (metrics) evaluations. They suppose that the productivity and defect detection rate are the popular metrics for measuring automation degree of MDD processes. Some quality goals such as well-establishment and precision are especially important in MDE [18 -20]. In [21], authors also developed open source tool aims to address quality measurement and prediction process to achieve automatically. In [22], authors presented the most recent challenges faced in the process to make model transformation more sophisticated. According to the literature, it is necessary to extend the assessment to other quality attribute of modifiability using object-oriented metrics. We conducted the comparative study for measuring the modifiability of MDD generated source code using object-oriented metrics. Therefore, we obtain more reliable findings.

3. Rule based model transformation framework

Model transformations become essential with the evolution of model driven development. It is a mechanism of automating the manipulation of models. A transformation is the automatic generation of a target model from source model using transformation definition. These transformations definition is a set of transformation rules that define how a model in the source language can transform into target language. These rules are descriptions of how one or more constructs in the source language can be transformed into one or more constructs in the target language.

In this section, we provide an overview of the framework and their underlying architecture. The proposed architecture is divided into three major parts corresponding to the main capabilities of the proposed framework. These components are parser, transformer and code generator. The parser receives an input model, written in Umple language, tokenizes it and passes it to the next component transformer. The transformer is a knowledge

based rule engines. It has received the tokens previously obtained and transforms them into internal representation consistent with target source code model using predefined set of Drools mapping rules. It is more correctly classified as a production rule system. It is a kind of Rule engine and also an Expert system, the validation and expression evaluation Rule Engines. It is declarative programming and allows to present what to do. The key advantage of rule engine is that using rules can make it easy to express solution to difficult problems and consequently have those solutions verified. The code generator translates the internal representation into target artifacts: source code as Java, XML and android activity class. Each component is tested independently to ensure that the input is processed correctly and the resulted output is validated [1]. Figure 1 describes the overall architecture of the proposed system. We have used Umple as input and transformed to mobile application. In this case of model transformation, we have applied Drool Rule based transformation.

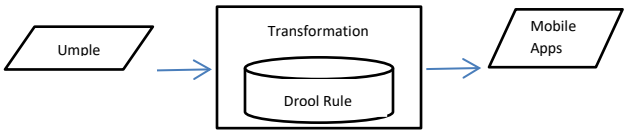


Figure 1 Overall Architecture of model transformation

3.1. Rule-based Inference System

In this system, we present a rule-based model driven approach to generate android application from text-based modeling language. Rule languages and inference engines incorporate reasoning capabilities are used in mobile application development system. A rule is made up of a collection of conditions associated with a sequence of actions to be applied to each collection of facts matching the rule condition. The proposed model transformation rules are based on Drools rule inference engine [2]. It is improved to reach the generation of mobile applications source code and introducing new concern in model driven mobile engineering. The core of the Drool suites and advanced Inference Engine using are improved Rete algorithm for object pattern matching. Rules are stored in the production memory, while facts are maintained in the working memory. The production memory remains the same during an analysis session, i. e, rules cannot be added or removed or changed. The contents of the working memory on the other hand can change. Facts may be modified, removed or added by executing rules or from external sources. After changing in the working memory, the inference engine is triggered and it works out which rules become "true" for the given facts. If there are multiple selected rules, their execution order will be managed via the Agenda, using a conflict resolution strategy.

3.2. Drools Transformation Rule

Drools rules are defined using Java-like language. It is a Business Logic integration Platform (BLiP). With the runtime, we create a working memory. The syntax of rule is shown as follows:

Rule
Rule <Rule Name>
When <Condition>
then <Action>

Rule: A rule is nothing but the logic that will be applied to incoming data. It has two main parts; when and then.

When: works out the condition on which Rule will be fired.

Then: the action; if the rules met the condition, they define what work this rule performs.

Step 1: Create a.drl (droolRule.drl)file where we will define the rules.

Step 2: Create Person POJO class.

The proposed rule engine consists three parts: umple2model, umple2view and umple2controller according to android model, view and controller perspective. Table 1 shows the sample form of Drools transformation rule for simple variable declaration for Account Title. Umple2Model.drl transforms incoming abstract syntax model (ASM) into plain java object (POJO). Umple2View.drl transforms ASM into android user interface XML file and Umple2Controller.drl transforms ASM into android activity class. The code generator receives the POJO model for model layer, XML model for view layer and android model for controller layer. The generator use the java development tool (JDT-core) to generate POJO class and android class source code. It is also used the JDOM to generate XML user interface file.

Table 1: Transformation Rule Sample

```
Umple          String AccountTitle;

Rule "VariableDeclaration"
Dialect "java"
when
$st : SyntaxTree(status==SyntaxTree.VAR_DECLARE)
then
TypeDeclaration type=AST2Android.Variable_Decl($st.getType());
CompilationUnit cu=$st.getCu();
$st.setStatus(SyntaxTree.ACTIVITY_CREATE);
$st.setType(type);
$st.setCu(cu);
update($st);
end
```

```
POJO (Model Layer) String AccountTitle;
PublicvoidsetAccountTitle(String at){
accountTitle=at;}
public String getAccountTitle(){
return accountTitle;}
}

XML (View Layer)
<EditText android:id=
"@+id/ txtaccountTitle"
android:layout_height=
"wrap_content"
android:layout_width=
"wrap_content"/>

Android (Controller Layer)
private EditText txtaccountTitle;
private String accountTitle;
txtaccount-
Title=(EditText)findViewById(R.id.accountTitle);
```

4. Experimental results and comparison

In model driven development, there are two important criteria to evaluate the quality of model transformation. They are transformability and modifiability. The consistency of source and target model and the assessment of transformability are evaluated in [1]. According to these results, we extend to evaluate the assessment of quality of modifiability in this model driven transformation.

4.1. Modifiability

Changes made for the requirements are rendered quality of the code in the models the code. This fact becomes challenges in quality of model driven development [quality]. To address these issues, this paper has proposed the evaluation of modifiability of model driven transformation using object oriented metrics. This modifiability is decomposed into traceability of model elements and well-designated or not being too complex. Moreover, the extent to which a model transformation can be adapted to provide different or additional functionality. The main reason for modifying a model transformation is changing requirements. Another reason is that the (domain specific) language in which the source and/ or target model is described which may be subject to changes. Modifiability captures the amount of effort need to modify a model transformation. It controls the visibility of system development. Such controls contribute to modularity, an essential aspect of software engineering that promotes software maintainability. In object-oriented programming, we note that these classes form the modules of programs. From the modularity perspective, modules should be as independent as possible with minimal coupling.

We have also performed a comparative study using both our proposed system and JUSE4 Android [8, 9] based on object oriented metrics. These metrics indicate quality of source code directly. We evaluate the quality of model driven for model transformation of generated source code quality and prior approach's generated source code quality. The results are used to evaluate a model is complete or suitable for automation or a modeling technique is appropriate for a target transformation. Therefore we have identified metrics to examine in this process.

4.2. Metrics

We describe the metrics for assessing the quality attributes for model transformation. Those metrics are applicable to language definition and characteristics of languages. For modifiability, we determine encapsulation, polymorphism and reusability as the quality criteria. These are described in table 2.

Table 2: Object oriented quality criteria

Quality of Criteria for Comparative Study				
No	Quality Criteria	Metrics	Acronyms	Desired Results
1	Encapsulation	Methods of hiding factor	MHF	High
		Attribute hiding factor	AHF	High
2	Polymorphism	Number of method overridden by a subclass	NMO	High
		Polymorphism factor	PF	High
3	Reusability	Reuse ratio	RP	High
		Specialization ratio	SR	High

A. Data Collection

To evaluate the modifiability of transformation, we have collected the metric values by using Eclipse metrics Plug-in [23]. It is an open source metrics calculation tools which measures various metrics and detects cycle in package and type dependencies. At first, we have generated the android application from different approaches of proposed and prior approach respectively. In the next step, we enable Eclipse Metrics Plug-in on each generate source code that give common solution. Finally, we extracted the mean, standard deviation and maximum metric values for each generate source code. In our comparative study, we have collected the average metric values from the proposed

and prior generated source code with respect to the quality criteria. Table 3 shows extracted metric values.

4.3. Encapsulation

It is the bringing together of a set of fields and methods into an object definition and hiding their internal working from the users of the object. By encapsulation, the way an object or its fields and methods are structured which is not visible to the users of the object. It is also facilitated by bundling and information hiding. It enhances the software maintainability. Encapsulation increases the cohesiveness of data and methods through bundling and reduces the strength of coupling between software components through information hiding. For encapsulation, we have measured method hiding factor (MHF) and attribute hiding factors (AHF) using the following equation 1 and 2. MHF and AHF are indicators to show how well methods and attributes are hidden inside classes. The results are presented in table 3. The comparisons of MHF values and AHF values of proposed system and prior systems are shown in the following figure 2 and 3 respectively. These metrics are measured at system level and high metric values are expected. The results are compared with prior approach and our proposed approach. This result means that we achieve the higher method hiding factor and attribute hiding factors.

Let V (M)= number of classes where the method M is visible, then

$$MHF= 1-\frac{\sum V(M)/(Total\ numbers\ of\ classes-1)}{Number\ of\ methods\ in\ all\ classes} \quad (1)$$

Let V (A) = number of classes where the attribute A is visible, then

$$AHF= 1-\frac{\sum V(A)/(Total\ numbers\ of\ classes-1)}{Number\ of\ attributrd\ in\ all\ classes} \quad (2)$$

By using the equation 1, we have calculated the MHF value. The result described that the ratio of number of classes where the visible method M is higher, the MHF value is lower.

By using the equation 2, we have calculated the AHF value. The result described that the ratio of number of classes where the visible attribute A is higher, the AHF value is lower.

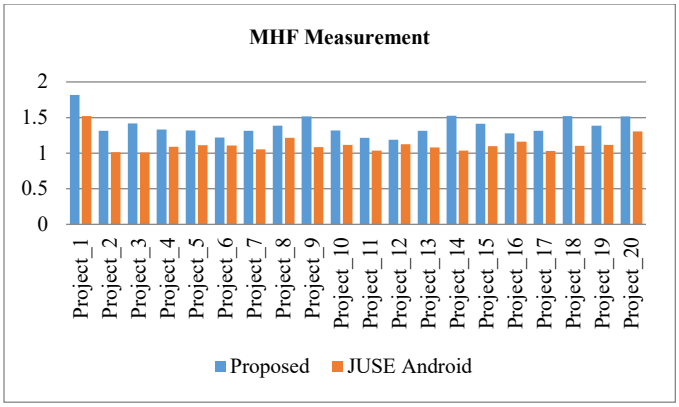


Figure 2 Comparison results of method of hiding factors

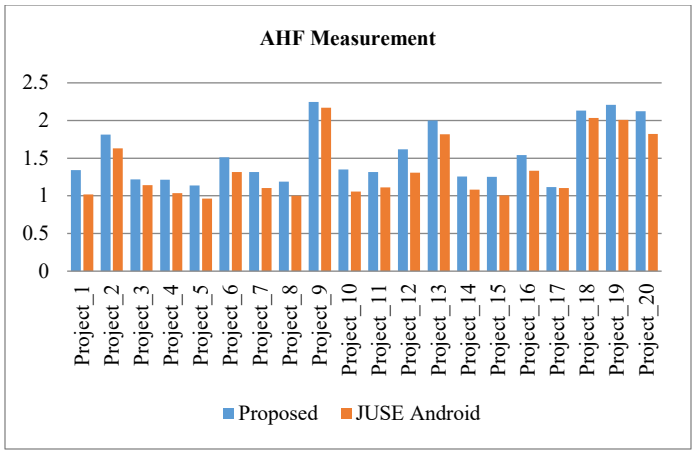


Figure 3 Comparison results of attribute hiding factors

4.4. Polymorphism

It is the ability of objects to respond to the same message but with the appropriate method based on their class definitions. For polymorphism, we have measured the number of method overridden by a sub class (NMO) and polymorphism factors (PF). Results are described in figure 4 and 5 respectively.

For NMO, we have determined the number of methods in a subclass overridden from its base class by using equation 3. Moreover, we determine the PF by using following equation 4. By using this equation 4, we present the PF value in table 3. To be specific, NMO is a class-level metric, which refers to the number of methods overridden by a single subclass, while PF is a system level metric, which measures the degree of method overriding in the whole type tree. These values are desired to be high. The results are compared with prior approach and our proposed approach. This result means that we achieve the higher number of methods overridden by a subclass and polymorphism factor.

NMO= the number of methods in a subclass overridden from its base class (3)

$$PF= \frac{\sum_{i=0}^{TC} M0(Ci)}{\sum_{i=0}^{TC}[(Mn\ Ci)x\ DC\ (Ci)]} \quad (4)$$

Where TC = the total number of classes

M_n (C_i)= Number of new methods of the class C_i

M_o (C_i)= Number of overriding methods of the class C_i

DC(C_i)= Number of Descendant of the class C_i

We have calculated the NMO value by using the number of methods in a subclass overridden from its base class and applying equation 3. They are presented in table 3.

We have calculated the PF value by using equation 4 and the results are described in table 3. The results present that number of overriding methods of the class is higher, the PF value is higher.

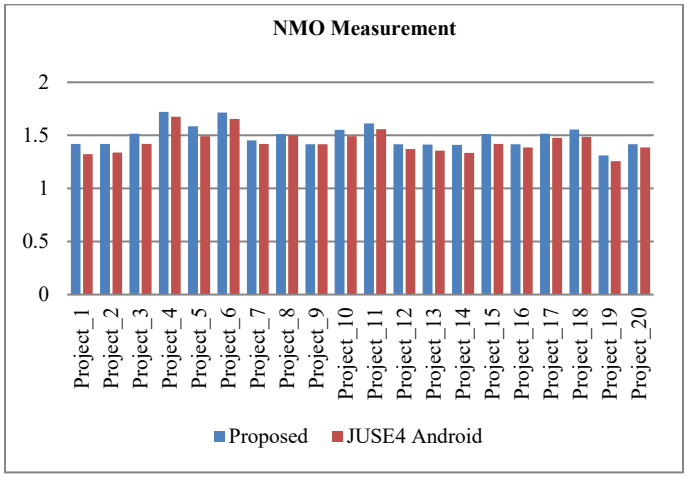


Figure 4 Comparison results of number of methods overridden by a sub class

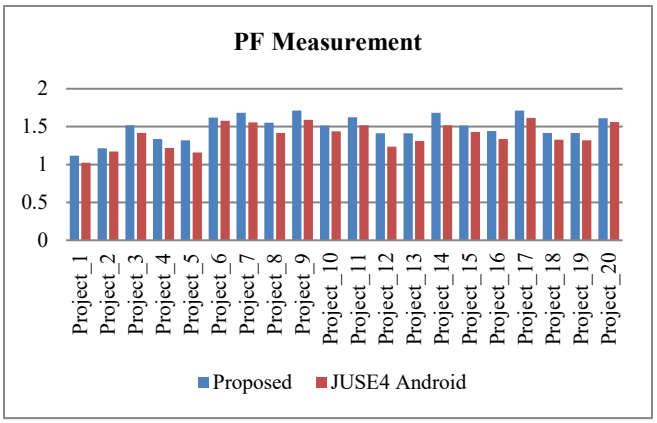


Figure 5 Comparison results of polymorphism factors

4.5. Reusability

Reusability is the extent to which a model transformation can be reused by other model transformations. It refers to as-is reuse. It is especially relevant for model transformations when a source model has to be transformed into different target models or vice versa. For reusability, we have measured reuse ratio (RR) and specialization ratio (SR) and results are presented in figure 6 and 7.

We have determined the RR and SR by using the following equation 5 and 6. RR and SR are both system level reusability metrics. They are calculated as the ratios of subclass to all classes and to super classes, respectively. The results are presented in table 3. We have expected to be highly reused, large reusability metrics values are desirable.

$$RR = \frac{(Total\ number\ of\ Super\ classes)}{(Total\ number\ of\ classes)} \quad (5)$$

$$SR = \frac{(Total\ number\ of\ Sub\ classes)}{(Total\ number\ of\ super\ classes)} \quad (6)$$

We have calculated the RR and SR values by applying equations 5 and 6 respectively. There are more subclasses, the higher the RR and SR values.

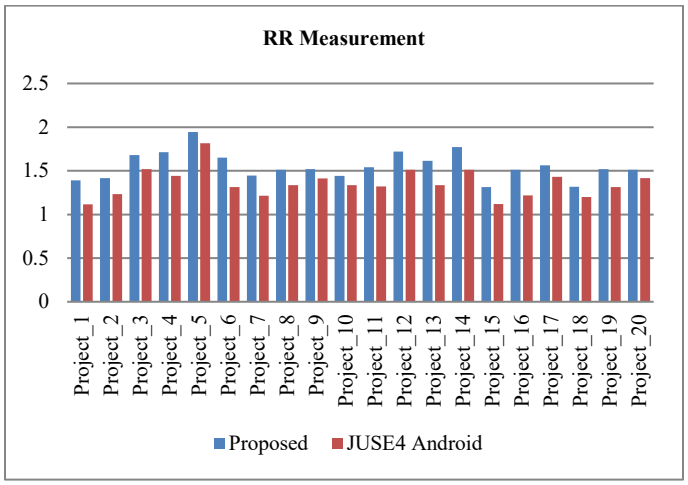


Figure 6 Comparison results of reuse ratios

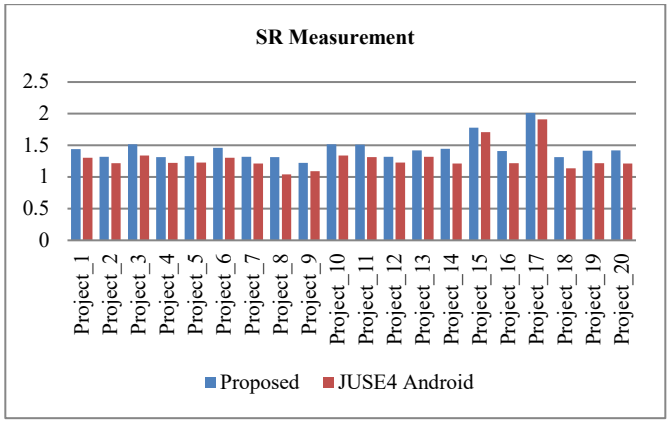


Figure 7 Comparison results of specialization ratios

5. Results and discussion

We have presented comparison results of our measurement. Expected results for our evaluation are shown in table 2. The larger a model transformation, the harder it is to understand and modify. Moreover, the number of signature and equations per function has a negative effect on consistency. If more similar signatures or equations have to be written, it is more likely that a different style is used.

The comparison results of our measurements are described in table 3. We have employed the result from MHF and AHF metrics for encapsulation, NMO and PF metrics for polymorphism and RR and SR for reusability to compare our approach and prior approach.

We have used private, protected and public keywords to control the accessibility to the method and attributes inside a class. According to these facts, we have planned quality attributes of our system to achieve higher modifiability.

Table 3: Metric values collected by Eclipse Metrics Plug in

Project	MHF		AHF		NMO		PF		RR		SR	
	Proposed	Prior	Proposed	Prior	Proposed	Prior	Proposed	Prior	Proposed	Prior	Proposed	Prior
Project_1	1.817	1.52	1.341	1.017	1.418	1.324	1.118	1.024	1.391	1.117	1.441	1.305
Project_2	1.312	1.012	1.814	1.631	1.418	1.339	1.214	1.172	1.415	1.234	1.318	1.216
Project_3	1.418	1.01	1.218	1.141	1.516	1.418	1.516	1.418	1.681	1.519	1.515	1.341
Project_4	1.332	1.091	1.216	1.037	1.721	1.675	1.338	1.219	1.712	1.441	1.312	1.225
Project_5	1.318	1.113	1.137	0.964	1.584	1.492	1.321	1.158	1.944	1.814	1.331	1.226
Project_6	1.218	1.108	1.512	1.314	1.714	1.654	1.62	1.578	1.651	1.315	1.461	1.306
Project_7	1.314	1.052	1.315	1.103	1.454	1.418	1.681	1.554	1.445	1.215	1.317	1.211
Project_8	1.386	1.216	1.189	1.003	1.512	1.5	1.551	1.418	1.512	1.335	1.312	1.041
Project_9	1.514	1.084	2.247	2.171	1.416	1.415	1.712	1.589	1.518	1.412	1.224	1.091
Project_10	1.317	1.117	1.351	1.056	1.551	1.491	1.512	1.438	1.441	1.337	1.516	1.338
Project_11	1.215	1.034	1.314	1.114	1.612	1.558	1.623	1.519	1.541	1.319	1.511	1.314
Project_12	1.188	1.127	1.618	1.306	1.416	1.371	1.412	1.237	1.721	1.512	1.317	1.227
Project_13	1.315	1.081	2.001	1.818	1.412	1.356	1.412	1.311	1.615	1.336	1.418	1.318
Project_14	1.523	1.034	1.258	1.084	1.411	1.336	1.681	1.518	1.771	1.512	1.446	1.215
Project_15	1.412	1.098	1.252	1.004	1.513	1.418	1.512	1.429	1.314	1.118	1.781	1.711
Project_16	1.278	1.161	1.541	1.331	1.417	1.387	1.441	1.337	1.512	1.217	1.412	1.219
Project_17	1.314	1.033	1.118	1.105	1.516	1.477	1.711	1.616	1.561	1.431	2.012	1.911
Project_18	1.521	1.102	2.132	2.034	1.554	1.486	1.417	1.327	1.318	1.201	1.316	1.138
Project_19	1.386	1.117	2.21	2.007	1.312	1.258	1.415	1.318	1.518	1.312	1.416	1.216
Project_20	1.517	1.305	2.124	1.823	1.416	1.387	1.612	1.559	1.512	1.416	1.418	1.213

By applying the experimental results, our approach has higher value than prior approach in encapsulation, polymorphism and reusability. Moreover, our expected result is that the high metric values are preferable.

These experimental results show how well methods and attributes are hidden inside classes. Therefore, our approach can help system develop methods and attributes which are hidden inside classes more efficiently. This means that our approach is more efficient than others. However, these results describes that they are a little bit higher than prior approach. By using these results, we will enhance our approach to achieve more efficiently and effectively model driven development process.

6. Conclusion

Model transformations become essential with the evolution of model driven development. It is an automatic generation of a target model from source model by using transformation definition. Modifiability is key issues in quality of this transformation. To address this issue, we have evaluated the modifiability of quality of model transformation using object oriented metric. This modifiability is decomposed into traceability of model elements and well-designated or not being too complex. Moreover, the extent to which a model transformation can be adapted to provide different or additional functionality. The main reason for modifiability of a model transformation is changing requirements. In this paper, we have performed the comparative study on our approach and prior approach to determine modifiability to develop more efficient mobile application system. We have determined the encapsulation, polymorphism and reusability as quality metrics. These metrics are measured at system level. We have used private, protected and public keywords to control the accessibility to the method and attributes inside a class. According to these facts, we have planned quality attributes of our system to achieve higher modifiability. The determination of experimental results represent that we achieve high score from comparison of our approach and prior approach. This means that our system is more traceability and

well-designated. Using these findings, we will enhance our approach to achieve higher efficiency and quality. In the future work, we will investigate more quality attribute for high accuracy of system development. Moreover, we will also evaluate the impact of transformation rules.

References

[1] Thu, E. E, Nwe N, "Model Driven Development of Mobile Applications Using Drools Knowledge-based rule" proceeding of SERA 2017, June 7-9, 2017, London, UK.

[2] De Lay, E, Jacobs D, "Rules-based Analysis with JBoss Drools: Adding Intelligence to Automation", ICALEPCS 2011, Proceeding of ICALEPCS Genoble, France

[3] Son, H.S, Kim, W.Y and Chul, R.Y, "MOF based Code Generation Method for Android Platform", International Journal of Software Engineering and Its application, Vol 7, No 3, Hongik University, Sehong Campus, Korea, 2013.

[4] Son, H. S, Kim, J.S, Chul, R. Y, SMTL Oriented Model Transformation Mechanism for Heterogeneous Smart Mobile Models, International Journal of Software Engineering and its Applications, Volume 7, No3, Sejong Campus, Korea, 2013.

[5] Parada, A.G, Lisane, B, A Model Driven Approach for Android Application Development, Brazilian, Symposium on Computing System Engineering (SBESC), Pelotas, Brazil, 2012.

[6] Parada, A.G.; Milena, R.S.; Automating mobile application development: UML-based code generation for Android and Window phone, Journal of Theoretical and Applied Information (RITA), volume 22, Pelotas, Brazil, 2015.

[7] Solheim, L and Neple, T, Model Quality in the Context of Model-Driven Development, Proceeding of 2nd International Workshop on Model- Driven Enterprise Information Systems (MDEIS'06), pp. 27-35, 2006.

[8] Silva, L.P, Abreu,F.B.: A Model-Driven Approach for Mobile Business Information Systems Applications, ACM/IEEE 17th International Conference on Model Driven Engineering Languages and Systems, MODELS 2014, QUASAR/ISTAR/ISCTE-IUL, Lisboa, Portugal.

[9] Silva, L.P, Abreu,F.B.: Model-Driven GUI Generation and Navigation for Android BIS Apps, MODELSWARD 2014, Portugal.

[10] Baowen, X, Di.W.: A metrics-based Comparative Study on Object-Oriented Programming Languages, 27th International Conference on Software Engineering and Knowledge Engineering, SEKE' 2015, USA.

[11] van Amstel, M.F; Lange, C.F.J.; van den Bran, M.G. J.: Metrics for analyzing the quality of model transformations, Proceeding s 12th ECOOP Workshop on Quantitative Approaches on Object Oriented Software Engineering (QAOOSE 08, Paphos, Cyprus, July 8, 2008.

[12] Rudiger, L, Jonas, L.: Comparing Software Metrics Tools, International Symposium on Software Testing and Analysis, ISSTA'2008, USA.

[13] Omar, E.B, Bragun, B, Automatic Code Generation by Model Transformation from Sequence Diagram of System's Internal Behavior International Journal of Information Technology, Hassan 1st University, Morocco, 2012.

[14] Steeg, C.C, Gotz, F, Model Driven, Data Management in Android with the Android Content Provider, https:// code.goodle.com/archieve/p/mdsd-android-content-provider/Bingen, Germany, 2011.

[15] https://github.com/umple/umple

[16] Mohagheghi, P, Aagedal, J, Evaluating Quality in Model-Driven Engineering, Proceeding of 29th International Conference on Software Engineering Workshops (ICSEW'07), 2007

[17] Mohagheghi,p.: An Approach for Empirical Evaluation of Model Driven Engineering in Multiple Dimensions, MODELPLEX, Oslo, Norway, 2010.

[18] Chidamber, S.R, Kamerer, C.F.: A metrics suite for object-oriented design, IEEE Transaction of Software Engineering, pp-(20)6, 1994: 476-493.

[19] Henderson-Sellers, B.: Object-oriented metrics: measures of complexity, Prentice Hall, 1995.

[20] Badreddin, O, Lethbridge, T.C, Forward, A, A Novel Approach to Versioning and Merging Model and Code Uniformaly, in MODELSWARD 2014, Proceeding of the 2nd International Conference on Model-Driven Engineering and Software Development, 2014.

[21] Kocaguneli, Ekrem, et al.: Prest: An Intelligent Software Metrics Extraction, Analysis and Defect Prediction Tool, 21th International Conference on Software Engineering and Knowledge Engineering, SEKE'2009, USA.

[22] Pallavi, K, Suita, P.U: Model Transformation, Concept, Current Trends and Challenges, International Journal of Computer Applications, Volume 119, No 14, Nasik, Manarashtra, India, 2015.

[23] http://metrics.sourceforge.net

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 252-260 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Constant Envelope DCT-based OFDM System with M-ary PAM Mapper over Fading Channels

Rayan Hamza Alsisi [*] , Raveendra Kolarramakrishna Rao

Faculty of Engineering, Electrical and Computer Engineering, The University of Western Ontario, London, ON, N6A 5B9, Canada

ARTICLEINFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 31 January, 2018

Keywords:
Bit Error Rate analysis
Discrete Cosine Transform
Constant Envelope signals

ABSTRACT

Constant Envelope Discrete Cosine Transform based Orthogonal Frequency Division Multiplexing (CE-DCT-OFDM) system with M-ary Pulse Amplitude Modulation (PAM) mapper is considered. In the system phase modulation is used to achieve constant envelope signals that have 0 dB Peak-to-Average-Power Ratio (PAPR). Transmission of such signals permit high power amplifiers in the system to operate with maximum power efficiency. The performance of CE-DCT-OFDM system is examined over Additive White Gaussian Noise (AWGN) and over fading channels. Closed-form expressions for Bit Error Rate (BER) over Ricean and Rayleigh channels are derived. The performances of CE-DCT-OFDM and conventional DCT-OFDM systems are compared as a function of Input power Back-Off (IBO) and Signal-to-Noise Ratio (SNR) for the Traveling-Wave Tube Amplifier (TWTA) model. Results show that CE-DCT-OFDM system offers superior BER performance compared to DCT-OFDM system and has other advantages as well.

1 Introduction

Fast Fourier Transform based Orthogonal frequency division multiplexing (FFT-OFDM) is widely adapted in a variety of communication standards due to its attractive properties such as high spectral efficiency and low complexity of the receiver, particularly, over multipath fading channels [2]. In an FFT-OFDM system, complex orthogonal exponential functions are used as basis functions. Instead, orthogonal cosinusoidal functions can be utilized as basis to create multicarrier system. Such a system utilizes Discrete Cosine Transform (DCT) [3] and is referred to as DCT-OFDM system. Several researchers have been investigating the use of DCT in OFDM system [4,5,6], as it has several advantages over conventional FFT-OFDM system. They are:

1. DCT is well known to have excellent spectral compaction and energy concentration properties. As a result, the channel estimation and also the system performance can be improved in noisy environments [7].
2. DCT is widely adopted in image/video coding standards (e.g. JPEG). Using IDCT for modulation and DCT for demodulation in an OFDM system, results in better integrated system design and reduced overall implementation cost are possible [8].
3. DCT uses real arithmetic compared to complex arithmetic in the case of FFT. This reduces signal processing complexity and power consumption, especially, when M-ary Pulse Amplitude Modulation (MPAM) mapper is used in DCT-OFDM system [7].
4. In the presence of frequency offset, due to the energy-compaction property of DCT, the inter-carrier interference (ICI) coefficients in DCT-OFDM system are concentrated around the main coefficient. As a result, DCT-OFDM system is robust to Carrier frequency offset (CFO) [7].
5. When MPAM mapper is used in DCT-OFDM system, it requires half of bandwidth required by an FFT-OFDM system, with the same number of subcarriers [9].

One of the major drawbacks of an OFDM system is

[*]Corresponding Author Name: Rayan Hamza Alsisi, Email: ralsisi@uwo.ca,
This paper is an extension of work originally presented in (SysCon) [1]

https://dx.doi.org/10.25046/aj030131

the high PAPR of transmitted signals in an FFT-OFDM system. When high PAPR signals are amplified using non-linear power amplifier, severe signal distortion will occur. Therefore, power amplifier with suitable power backoff is required in the system. Without appropriate power backoff, the system suffers from spectral broadening, intermodulation distortion, and, consequently, performance degradation. The problem can be mitigated by increasing the power backoff, but this results in poor power efficiency. In mobile devices with limited battery supply power efficiency is required to be as high as possible [10]. Several techniques have been suggested to mitigate the problem of high PAPR in an OFDM system such as coding, partial transmission sequences, clipping, tone reservation, and filtering [11,12,13]. These techniques offer a variety of trade-offs in terms of complexity, performance and spectral efficiency.

An alternative approach to completely eliminate the PAPR problem in an OFDM system is based on signal transformation. In this technique, signal transformation occurs at the transmitter prior to modulation and an inverse transformation at the receiver prior to demodulation. In [14,15,16,17] phase modulation and demodulation are considered in OFDM systems. Such systems are characterized by constant envelope signal with 0 dB PAPR, and hence suitable for power amplification close to the saturation level of nonlinear power amplifier. While FFT-OFDM systems with phase modulation have been extensively studied in the literature, DCT-OFDM system with phase modulation has not received much attention. In this paper, therefore, DCT-OFDM system with phase modulation referred to as CE-DCT-OFDM is presented and examined. The intent of this paper is to present a generalized model of CE-DCT-OFDM system that can be used to examine its performance. The BER analysis of CE-DCT-OFDM system in AWGN channel is presented and then the analysis is extended to the case of fading channel, as over practical communication channels signal fading is always present.

This paper is organized as follows. Section II describes the generation of DCT-OFDM signal. Section III introduces phase modulation in DCT-OFDM system. CE-DCT-OFDM system with MPAM mapper is described in Section IV, and its performance is analysed in AWGN channel. Section V deals with performance analysis of CE-DCT-OFDM system over fading channels. Finally, the paper is concluded in Section VI.

2 Baseband DCT-OFDM Signal

The process of generating DCT-OFDM signal is shown in Figure 1. The signal can be represented by

$$f(t) = \sum_{n=0}^{N-1} C_n \, \varphi_n(t), 0 \le t < T, \tag{1}$$

$$\varphi_n(t) = \begin{cases} \sqrt{\frac{2}{T}} cos 2\pi f_n t, & 0 \le t < T, \\ 0, & otherwise \end{cases} \tag{2}$$

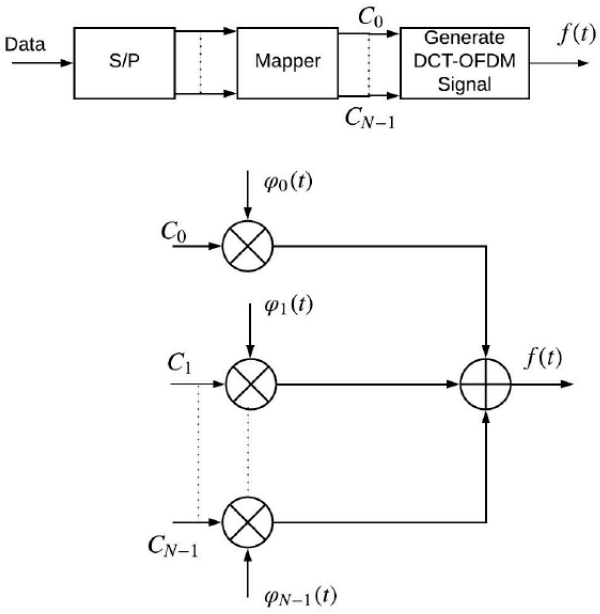


Figure 1: Block diagram of DCT-OFDM signal generator

The cosinusoidal function $\varphi_n(t)$ is the n^{th} orthogonal signal with frequency $f_n = n/2T$ and repersents the n^{th} subcarrier. The subcarrier spacing is $1/2T$. The subcarriers are orthonormal over $0 \le t \le T = NT_s$. That is,

$$\int_0^T \varphi_n(t)\varphi_k(t)dt \tag{3}$$

$$= \int_0^T \sqrt{\frac{2}{T}} cos 2\pi f_n t * \sqrt{\frac{2}{T}} cos 2\pi f_k t \, dt \tag{4}$$

$$= \begin{cases} 1 & n = k \\ 0 & otherwise \end{cases} \tag{5}$$

The time duration of the OFDM symbol is $T = NT_s$; T_s is the time duration of DCT-OFDM symbol. It is noted that $T_s = kT_b$ and $M = 2^k$. T_b denotes the bit duration. $C_n (n = 0, 1, .., N-1)$ are N independent data symbols obtained from MPAM signal constellation. The DCT-OFDM signal can thus be represented by

$$f(t) = \sqrt{\frac{2}{T}} \sum_{n=0}^{N-1} C_n \cos \pi nt/T, 0 \le t < T, \tag{6}$$

3 DCT-OFDM Signal with Phase Modulation

The phase modulated bandpass signal can be represented by

$$s(t) = A_c cos(2\pi f_c t + \phi(t)) \tag{7}$$

Where A_c and f_c are the carrier amplitude and frequency. The phase in (7) is proportional to $f(t)$ and is given by

$$\phi(t) = h_p f(t) \tag{8}$$

where h_p is the modulation index.

In DCT-OFDM system, $f(t)$ is real for MPAM mapper [9]. The advantage of DCT-OFDM system with phase modulation is that the transmitted signals have peak and average powers the same and hence, their PAPR is 0 dB. Figure 2 shows a comparison of instantaneous powers of DCT-OFDM and CE-DCT-OFDM signals.

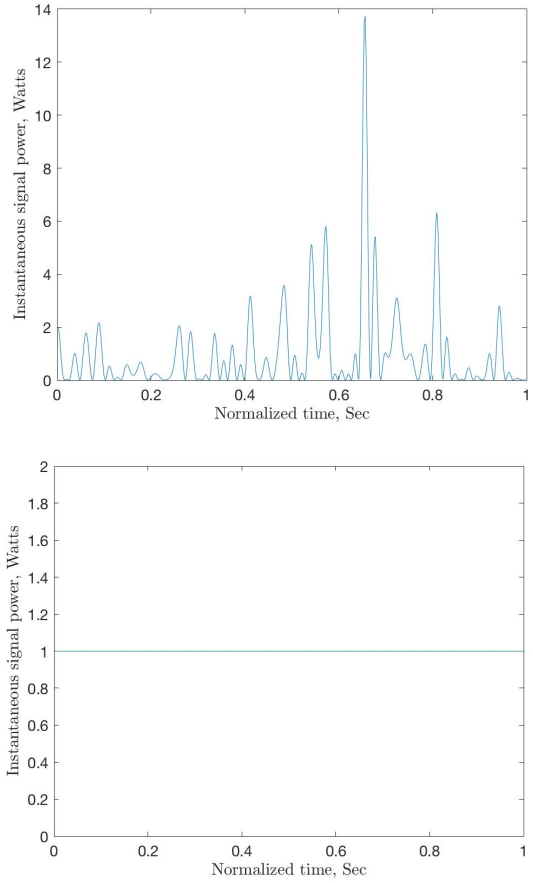


Figure 2: Instantaneous signal power: (a) DCT-OFDM signal and (b) CE-DCT-OFDM signal.

4 CE-DCT-OFDM Transmitter

The block diagram of CE-DCT-OFDM transmitter is shown in Figure 3. The output of the system can be written as:

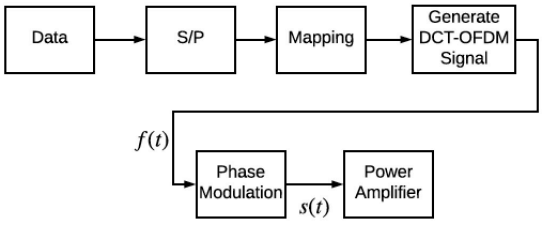


Figure 3: Block diagram of CE-DCT-OFDM transmitter

$$s(t) = A_c cos\left(2\pi f_c t + \sqrt{\frac{2}{T\sigma_s^2}} h_p \sum_{n=0}^{N-1} C_n cos(\pi nt/T)\right) \quad (9)$$

where $0 \le t \le T$, A_c is the signal amplitude and f_c is the carrier frequency. $\sigma_s^2 = (M^2-1)/3$ is the variance of the data symbols [16]. $\{C_n\}$ are MPAM data symbols, $C_n \in \{\pm 1, \pm 3, \cdots, \pm(M-1)\}$ for all n. The message signal is given by: $f(t) = \sqrt{2/T\sigma_s^2} \sum_{n=0}^{N-1} C_n cos(\pi nt/T)$. The average power of $s(t)$ is $P_s \left(= \int_0^T s(t)^2 dt/T = A_c^2/2\right)$ and the signal energy is $E_s = P_s T = A_c^2 T/2$. For k bits of information per symbol per transmission, the average bit energy is $E_b = A_c^2 T/2N \log_2 M = A_c^2 T/2Nk$.

4.1 Bandwidth Considerations

Phase modulated signals are complex to analyze for their bandwidth. However, simple observations can be used to get rough idea about the bandwidth of CE-DCT-OFDM signals. Using Maclaurin series, the PM signal described in equation (7) can be written as

$$s(t) = A_c(cos2\pi f_c t - h_p f(t) sin2\pi f_c t - \frac{h_p^2}{2!} f^2(t) cos2\pi f_c t + \cdots) \quad (10)$$

When h_p is small, the first two terms in the series are sufficient to represent CE-DCT-OFDM signal. That is,

$$s(t) \approx A_c cos2\pi f_c t - A_c h_p f(t) sin2\pi f_c t \quad (11)$$

This represents the narrowband case and the bandwidth of the signal is at least $2W$, where W is the bandwidth of $f(t)$. As h_p becomes larger, the bandwidth of the signal broadens. A useful expression for bandwidth of the signal is given by the root-mean-square (RMS) bandwidth [18] which is equal to $max(2h_p, 2)WHz$. The bandwidth of the message signal $f(t)$ is $W = (N/2T)Hz$

4.2 BER Analysis over AWGN Channel

The CE-DCT-OFDM receiver consists of a phase demodulator followed by the standard DCT-OFDM demodulator to recover the transmitted data symbols as shown in Figure 4. Each block will be analyzed below.

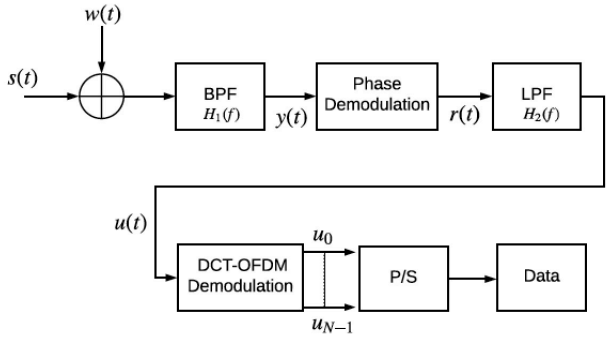


Figure 4: Block diagram of CE-DCT-OFDM receiver

4.2.1 Phase Demodulation

The noise $w(t)$ is modelled as additive white Gaussian with zero mean and power spectral density $N_0/2$. The received signal $s(t)+w(t)$ is fed to a BPF with transfer function $H_1(f)$ shown in Figure 5. The filter has a center frequency f_c and bandwidth B. It is noted that only a negligible amount of input signal power lies outside the frequency band $f_c - B/2 \leq |f| \leq f_c + B/2$. The bandwidth B is in excess of twice the message bandwidth W by an amount that depends on the deviation ratio of the signal $s(t)$. Thus, it is noted that the BPF allows the CE-DCT-OFDM signal without any distortion. The filtered narrow band noise $n(t)$ can be represented as

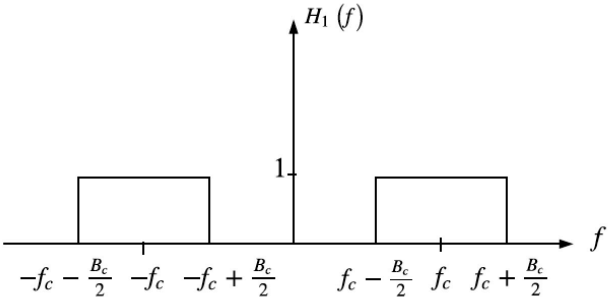


Figure 5: Ideal bandpass filter characteristic

$$n(t) = n_I(t)cos2\pi f_c t - n_Q(t)sin2\pi f_c t \quad (12)$$

where $n_I(t)$ and $n_Q(t)$ are the in-phase and quadrature components of the zero-mean filtered Gaussian noise. $n(t)$ can also be represented as

$$n(t) = x(t)cos[2\pi f_c t + \Psi] \quad (13)$$

where

$$x(t) = \sqrt{[n_I(t)^2 + n_Q(t)^2]} \quad (14)$$

and

$$\Psi = tan^{-1}(n_Q(t)/n_i(t)) \quad (15)$$

The bandpass filter output $y(t)$ can be written as:

$$y(t) = A_c cos[2\pi f_c t + \phi(t)] + x(t)cos[2\pi f_c t + \Psi(t)] \quad (16)$$

Equation (16) in polar form is given by:

$$y(t) = v(t)cos[2\pi f_c t + \theta(t)] \quad (17)$$

where $v(t)$ represents the envelope and $\theta(t)$ is the phase angle which can be written as [18,19]

$$\theta(t) = \phi(t) + \varepsilon(t) \quad (18)$$

where

$$\varepsilon(t) = tan^{-1}\left\{\frac{x(t)sin[\Psi(t)+\phi(t)]}{A_c + x(t)cos[\Psi(t)+\phi(t)]}\right\} \quad (19)$$

is the noise signal. With the assumption of a high Carrier-to-Noise Ratio (CNR), $A_c >> x(t)$, equation (18) becomes

$$\theta(t) \approx \phi(t) + \frac{x(t)}{A_c} sin[\Psi(t) - \phi(t)] \quad (20)$$

The output of the phase demodulator is given by

$$r(t) = k_p \theta(t) \quad (21)$$

where k_p is the gain constant. With the large CNR assumption and choosing $k_p = 1/h_p$, equation (21) becomes.

$$r(t) = f(t) + n_d(t) \quad (22)$$

where

$$n_d(t) = \frac{k_p x(t)}{A_c} sin[\Psi(t) + \phi(t)] \quad (23)$$

The power spectral density $S_{Nd}(f)$ of $n_d(t)$ is related to the power spectral density $S_{NQ}(f)$ of $n_Q(t)$. That is [19,20],

$$S_{Nd}(f) = \left\{\frac{k_p}{A_c}\right\}^2 S_{NQ}(f) \quad (24)$$

where

$$S_{NQ}(f) = \begin{cases} N_0, & |f| \leq \frac{B}{2} \\ 0, & otherwise \end{cases} \quad (25)$$

The phase demodulator output is applied to a low pass filter of bandwidth equal to message bandwidth W. It is used to pass the message signal and reject out-of-band noise from $n_d(t)$. The ideal transfer function of the filter is

$$H_2(f) = \begin{cases} 1, & |f| \leq W \\ 0, & otherwise \end{cases} \quad (26)$$

The output of the low pass filter can be written as

$$u(t) = f(t) + n_u(t) \quad (27)$$

The power spectral density $S_{Nu}(f)$ of output noise $n_u(t)$ at the output of low pass filter is given by

$$S_{Nu}(f) = \begin{cases} N_0 k_p^2 / A_c^2, & |f| \leq W \\ 0, & otherwise \end{cases} \quad (28)$$

and the average output noise power is

$$\int_{-W}^{W} \frac{N_0 k_p^2}{A_c} df = \frac{2WN_0 k_p^2}{A_c^2} \quad (29)$$

4.2.2 DCT-OFDM Receiver

The DCT-OFDM receiver, as depicted in Figure 4, is composed of two stages: a demodulator and a detector. The demodulator projects the incoming signal using orthonormal bases and generates a vector whilst the detector applies detection algorithm to estimate the transmitted information symbols.

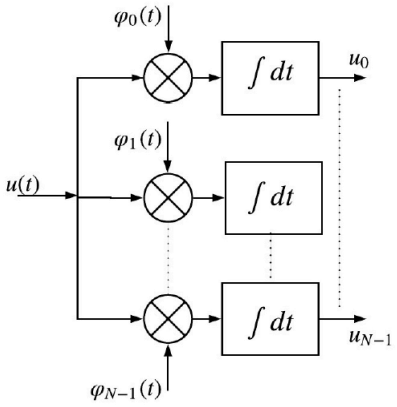


Figure 6: DCT-OFDM signal demodulation using correlators

The input to the bank of correlators (DCT-OFDM demodulator) in Figure 6 is the signal $u(t)$. The output of the demodulator is the vector $\bar{u}$. The j-th element of $\bar{u}$, can be expressed as:

$$u_j = \int_0^T u(t)\varphi_j(t)dt \tag{30}$$

$$= \int_0^T [f(t) + n_u(t)]\varphi_j(t)dt \tag{31}$$

$$= C_j/\sqrt{\sigma_s^2} + N_{uj}; j = 1, 2, .., N-1 \tag{32}$$

The mean of u_j is,

$$E[u_j] = E[C_j/\sqrt{\sigma_s^2} + N_{uj}] = E[C_j/\sqrt{\sigma_s^2}] = C_j/\sqrt{\sigma_s^2} \tag{33}$$

where $C_j \in \{\pm 1, \pm 3, ..., \pm(M-1)\}$. The mean is independent of the noise. However, the variance of u_j is dependent on noise and is given by:

$$Var[u_j] = \sigma_{uj}^2 \tag{34}$$

$$= E[(u_j - C_j)^2] \tag{35}$$

$$= E[(N_{uj})^2] \tag{36}$$

$$= E\left[\int_0^T n_u(t)\varphi_j(t)dt \int_0^T n_u(z)\varphi_j(z)dz\right] \tag{37}$$

$$= E\left[\int_0^T \int_0^T \varphi_j(t)\varphi_j(z) \cdot n_u(t)n_u(z)dtdz\right] \tag{38}$$

Taking the expectation operation inside the integrals, we can write

$$\sigma_{uj}^2 = \int_0^T \int_0^T \varphi_j(t)\varphi_j(z)E\left[n_u(t) \cdot n_u(z)\right]dtdz \tag{39}$$

$$= \int_0^T \int_0^T \varphi_j(t)\varphi_j(z)R_n(t,z)dtdz \tag{40}$$

where $R_{nu}(t,z)$ is the autocorrelation function of the noise process. The variance of u_j can be shown to be given by:

$$\sigma_j^2 = \frac{2WN_0k_p^2}{A_c^2}\int_0^T \varphi_j^2(t)dt \tag{41}$$

$$= \frac{2WN_0k_p^2}{A_c^2} \tag{42}$$

4.2.3 Probability of Bit Error

The symbol error rate (SER) can be shown to be given by [21]:

$$SER = \frac{M-1}{M}2P\left(N_{uj} > \frac{1}{\sqrt{\sigma_s^2}}\right) \tag{43}$$

$$= 2\frac{M-1}{M}\int_{1/\sqrt{\sigma_s^2}}^{\infty} \frac{1}{\sqrt{2\pi\left(2WN_0/A_c^2h_p^2\right)}} \tag{44}$$

$$e^{\left(-x^2/\left[2\left(2WN_0/A_c^2h_p^2\right)\right]\right)}dx$$

$$= 2\frac{M-1}{M}\int_{1/[2WN_0\sigma_s^2/A_c^2h_p^2]^{0.5}}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \tag{45}$$

$$= 2\left(\frac{M-1}{M}\right)Q\left(\sqrt{\frac{6h_p^2\log_2(M)E_b}{(M^2-1)N_0}}\right) \tag{46}$$

It is noted that for $h_p = 1$, (46) represents the SER for MPAM system [21]. For high CNR, the only significant symbol errors are those that occur at adjacent signal levels. The BER of CE-DCT-OFDM system, thus, can be approximated as [21]

$$BER \approx \frac{SER}{\log_2(M)}$$

$$\approx 2\left(\frac{M-1}{M\log_2(M)}\right)Q\left(\sqrt{\frac{6h_p^2\log_2(M)E_b}{(M^2-1)N_0}}\right) \tag{47}$$

The BER performance given by (47) is a function of E_b/N_0, signal-to-noise ratio, h_p, modulation index, and M, the number of amplitude levels in the MPAM mapper in the CE-DCT-OFDM system. The performance of CE-DCT-OFDM system with $h_p = 0.7$ for various values of M are illustrated in Figure 7, which shows that BER increases as M increases for fixed value of modulation index. For example at BER = 10^{-5} and $h_p = 0.7$, the SNR required for $M = 16$ is 14 dB more than that required for $M = 4$.

Figure 8 depicts BER performance of CE-DCT-OFDM system for 16-PAM mapper for $h_p = 0.3, 0.7$ and 1.2. It is observed that BER decreases as h_p increases for a fixed value of SNR. For example at BER = 10^{-5} the SNR required for $h_p = 0.3$ is 12 dB more than that required for $h_p = 1.2$. The BER performance of the system can be controlled by varying h and M as shown in Figure 9. For example, the system with $M = 16$ and $h_p = 1.7$ outperforms the system with $M = 4$ and $h_p = 0.2$ by nearly 9 dB at BER = 10^{-5}.

Figure 10 compares simulation results to theoretical BER given by (47) for CE-DCT-OFDM system using $N = 64$ subcarriers, and $M = 4$. For E_b/N_0 greater

than 15 dB, and for a small modulation index $h_p = 0.1$, simulation result is nearly the same as theoretical result. For large modulation index, for example $h_p = 0.8$, the theoretical BER is not as accurate as simulation result but still is within 1 dB of the former.

For the traveling-wave tube amplifier (TWTA) model, BER performance of CE-DCT-OFDM system is compared with that of DCT-OFDM system. The undesirable effects of TWTA nonlinearities can be reduced by increasing the input power backoff (IBO). For a given DCT-OFDM signal, we need to adjust the average input power so that the peaks of the signal are rarely clipped. That is, we will have to apply an IBO to the signal prior to amplification. Computer simulations are used to study the performance of the systems using nonlinear TWTA with various IBO levels. Figure 11 compares BER performance of 64 subcarrier 8-PAM CE-DCT-OFDM system and 8-PSK DCT-OFDM system with TWTA using IBO of 0 dB, 8 dB and 12 dB. At high SNR, CE-DCT-OFDM system provides significant performance improvement due primarily to the 0 dB backoff. The DCT-OFDM system with 0 dB IBO has an error floor at BER of 0.09. At the BER 10^{-3}, the IBO that results in the best DCT-OFDM system performance is 12 dB, with $E_b/N_0 = 16$ dB. However, the CE-DCT-OFDM system achieves this BER= 10^{-3} with $E_b/N_0 = 12$ dB which implies an advantage of 4 dB.

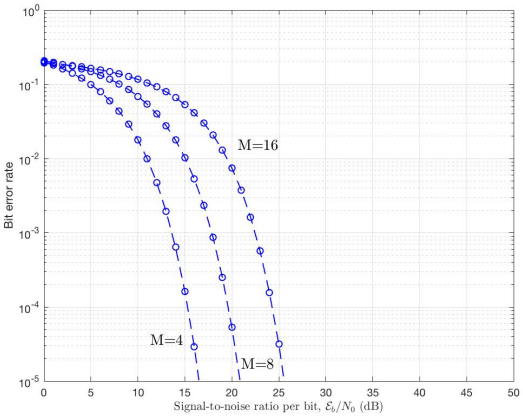


Figure 7: BER performance of CE-DCT-OFDM system over AWGN channel, as a function of M for $h_p = 0.7$

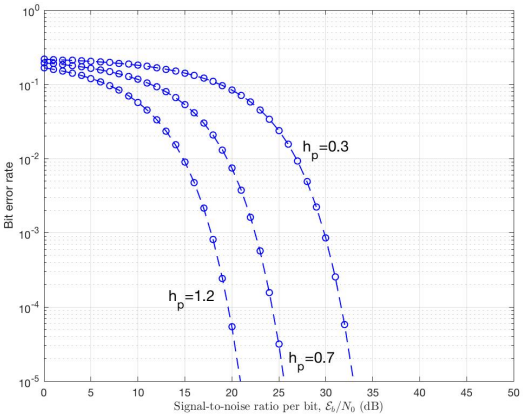


Figure 8: BER performance of CE-DCT-OFDM system over AWGN channel, as a function of h_p for $M = 16$.

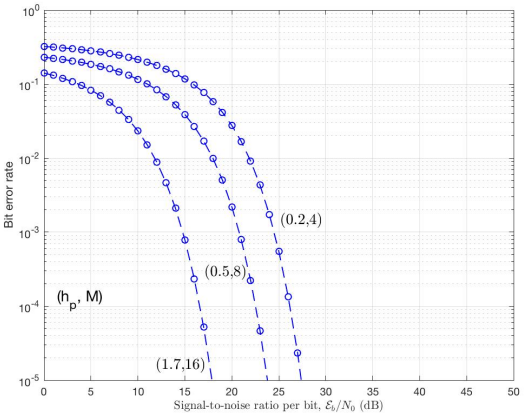


Figure 9: BER performance of CE-DCT-OFDM system over AWGN channel, as a function of (h_p, M).

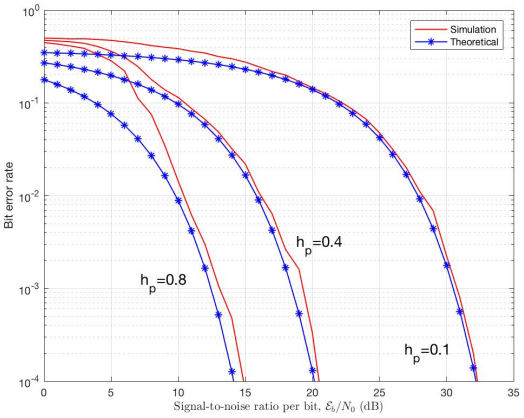


Figure 10: BER performance of 64 subcarrier CE-DCT-OFDM system with $M = 4$ over AWGN channel, as a function of h_p.

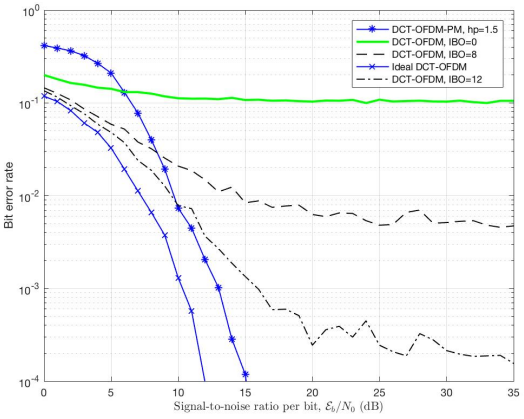


Figure 11: Comparison of 64 subcarrier 8-PAM CE-DCT-OFDM and DCT-OFDM systems for model TWTA for various of IBO.

5 Performance over Fading Channels

The received CE-DCT-OFDM signal over a fading channel can be expressed as:

$$r(t) = h(t) * s(t) + n(t) \tag{48}$$

where $h(t)$ represents the impulse response of the fading channel given by $h(t) = \alpha\delta(t)$. The instantaneous

SNR per bit and the average SNR per bit can be represented as $\gamma = \alpha^2 E_b/N_0$ and $\bar{\gamma} = E\left\{\alpha^2\right\} E_b/N_0$, respectively. To obtain the bit error rate (P_b) of CE–DCT-OFDM system over such a fading channel, the conditional BER is averaged over the Probability Density Function (PDF) of γ and can be written as [22]:

$$P_b = \int_0^\infty P(\gamma) p_\gamma(\gamma) d\gamma \tag{49}$$

where $P_b(\gamma)$ is given by:

$$P_b(\gamma) = 2\left(\frac{M-1}{M\log_2(M)}\right) Q\left(\sqrt{D\gamma}\right) \tag{50}$$

where $D = \frac{6h_p^2 \log_2(M)}{M^2-1}$. It is noted that $Q(z)$ in (50) is the well-known Q-function and it can also be written as:

$$Q(z) = \frac{1}{\pi}\int_0^{\pi/2} exp\left(-\frac{z^2}{2sin^2(\theta)}\right) d\theta \tag{51}$$

5.1 Rayleigh Fading Channel

For Rayleigh fading channel, the PDF of γ is given by [22]:

$$p_\gamma(\gamma) = \frac{1}{\bar{\gamma}} exp\left(\frac{\gamma}{\bar{\gamma}}\right), \gamma \geq 0 \tag{52}$$

Using (50)-(52) in (49), the average BER over Rayleigh fading channel can be shown to be given by

$$P_b = 2\left(\frac{M-1}{M\log_2(M)}\right)\frac{1}{\pi\bar{\gamma}}\int_0^{\pi/2}\int_0^\infty exp\left(-\frac{D\gamma}{2sin^2(\theta)} - \frac{\gamma}{\bar{\gamma}}\right) d\gamma d\theta \tag{53}$$

Upon simplification (53) [23], we get:

$$P_b = \left(\frac{M-1}{M\log_2(M)}\right)\left[1 - \sqrt{\frac{D\bar{\gamma}/2}{1+D\bar{\gamma}/2}}\right] \tag{54}$$

The BER given by (54) for CE-DCT-OFDM system over Rayleigh fading channel is a function of h_p, modulation index, M, number of levels in MPAM, and E_b/N_0, signal-to-noise ratio. The BER of CE-DCT-OFDM system for $M = 4$ is plotted as a function of h_p, and E_b/N_0 as shown in Figure 12. It is observed that BER increases as h_p decreases for a fixed value of SNR. For example at BER = 10^{-5} the SNR required for $h_p = 0.7$ is 7 dB more than that required for $h_p = 1.5$.

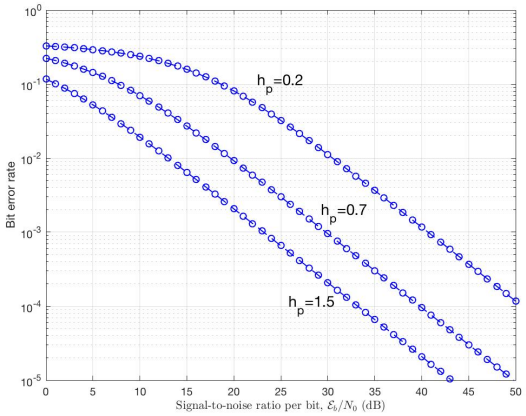


Figure 12: BER performance of CE-DCT-OFDM system over Rayleigh fading channel, as a function of h_p for $M = 4$.

5.2 Ricean Fading Channel

For the Ricean fading channel, the PDF of γ is given by [22]:

$$p_\gamma(\gamma) = \frac{(1+K)e^{-K}}{\bar{\gamma}} exp\left[-\frac{(1+K)\gamma}{\bar{\gamma}}\right] I_0\left[2\sqrt{\frac{(K+K^2)\gamma}{\bar{\gamma}}}\right], \gamma \geq 0 \tag{55}$$

Using (50), (51) and (55) in (49), P_b can be written as:

$$P_b = \frac{2}{\pi}\left(\frac{M-1}{M\log_2(M)}\right)\frac{(1+K)e^{-K}}{\bar{\gamma}}\int_0^{\pi/2}\int_0^\infty exp\left[-\frac{D\gamma}{2sin^2(\theta)} - \frac{(1+K)\gamma}{\bar{\gamma}}\right] I_0\left[2\sqrt{\frac{(K+K^2)\gamma}{\bar{\gamma}}}\right] d\gamma d\theta \tag{56}$$

Integrating (56) [23], BER of CE-DCT-OFDM system over Ricean fading channel can be written as:

$$P_b = \frac{2}{\pi}\left(\frac{M-1}{M\log_2(M)}\right)\int_0^{\pi/2}\frac{(1+K)sin^2(\theta)}{(1+K)sin^2(\theta)+D\bar{\gamma}/2} exp\left[-\frac{KD\bar{\gamma}/2}{(1+K)sin^2(\theta)+D\bar{\gamma}/2}\right] d\theta \tag{57}$$

The BER given by (57) is a function of h_p, M, K, Rice distribution parameter and E_b/N_0. The BER performance $K = 7$ dB and $M = 4$ is illustrated in Figure 13, are a function of h_p. It observed that there is improvement in BER as h_p increases, for example at BER = 10^{-5} the SNR required for $h_p = 1.5$ is 17 dB less than that required for $h_p = 0.2$.

The effect of parameter K on BER performance is illustrated in Figure 14, for $M = 4$ and $h_p = 0.5$. It is noted that the BER decreases as K increases, for example at BER = 10^{-5} and $h_p = 0.5$ the SNR required for $K = 2$ dB is 30 dB more than that required for $K = 18$ dB.

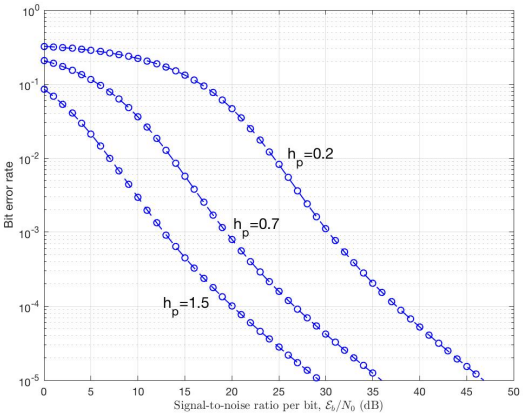


Figure 13: BER performance of CE-DCT-OFDM system over Ricean fading channel, as a function of h_p for $M = 4$.

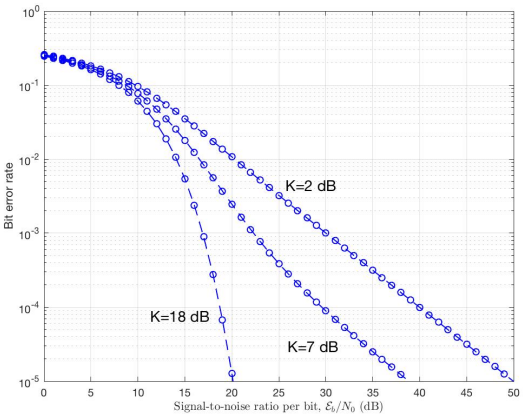


Figure 14: BER performance of CE-DCT-OFDM system over Ricean fading channel, as a function of K for $M = 4$ and $h_p = 0.5$.

6 Conclusions

A generalized description of CE-DCT-OFDM system is presented. In this system, phase modulation is used to eliminate the problem of PAPR. BER analyses of this system over AWGN and flat fading channels are presented and closed-form expressions for BER have been obtained. Improved BER performance is seen with increased value of modulation index at fixed transmission power. It is also observed that BER performance can be controlled by varying h_p and M as well. Simulation performance of CE-DCT-OFDM system over AWGN channel is also presented and compared with theoretical results. The results show that for a small modulation index, simulation result is nearly the same as theoretical result. With nonlinear the results show that CE-DCT-OFDM system has better BER performance than the conventional DCT-OFDM system when TWTA amplifier is used.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The first author would like to gratefully thank the Islamic University of Madinah and the ministry of higher education in Saudi Arabia for their support and scholarship.

References

1. Rayan H. Alsisi, Raveendra K. Rao, "Performance of constant envelope DCT based OFDM system with M-ary PAM mapper in AWGN channel," IEEE International Systems Conference (SysCon), pp.1-7, Montreal, April 2017. http://ieeexplore.ieee.org/document/7934716/
2. S. H. Han and J. H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarricer transmission," IEEE Wireless Commun., vol.12, no.2, pp.56-65, April 2005. http://ieeexplore.ieee.org/document/1421929/
3. Gi Hyun Kim, Honey Durga Tiwari, Chan Mo Kim, Yong Beom Cho, and Younggoo Kwon, "Implementation of DCT based OFDM system," International SoC Design Conference, IEEE, vol.2, pp.41-44, Busan, November 2008. http://ieeexplore.ieee.org/document/4815679/
4. C. Tang and D. Mandyam,"Performance of OFDM modem with alternative basis functions, " IEEE Radio and Wireless Symposium, pp.551-554, San Diego, October 2006. http://ieeexplore.ieee.org/document/1615216/
5. G. Mandyam, "Sinusoidal transforms in OFDM systems, " IEEE Trans. On Broadcasting, vol.50, no.2, pp.172-184, June 2004. http://ieeexplore.ieee.org/document/1304952/
6. R. Merched, "On OFDM and single carrier frequency domain systems based on trigonometric transforms," IEEE Signal Processing Letters, vol.13, no.8, pp.473-476, August 2006. http://ieeexplore.ieee.org/document/1658060/
7. N. Al-Dhahir, H. Minn, and S. Satish, "Optimum DCT-based multicarrier transceivers for frequency-selective channels," IEEE Trans. Commun., vol.54, no.5, pp.911-921, May 2006. http://ieeexplore.ieee.org/document/1632105/
8. S. Satish, N. Al-Dhahir, and H. Minn, "A DCT-Based Broadband Multicarrier Transceiver,", SoutheastCon, Proceedings of the IEEE, pp.175-180, Memphis, 2006. http://ieeexplore.ieee.org/document/1629345/
9. J. Tan and G. L. Stuber, "Constant Envelope Multi-Carrier Modulation," in Proceedings of IEEE Military Communications Conference, vol.1, pp.607-611, Anaheim, October 2002. http://ieeexplore.ieee.org/document/1180513/
10. Y. Tsai and G. Zhang, "Orthogonal Frequency Division Multiplexing with Phase Modulation and Constant Envelope Design," in Proc. of IEEE Milcom, pp.2658-2664, Atlantic City, October 2005. http://ieeexplore.ieee.org/document/1606068/
11. T.A. Wilkinson, A.E. Jones, "Minimization of the Peak to Mean Envelope Power Ratio of Multicarrier Transmission Schemes by Block Coding," IEEE VTC, pp.825-829, Chicago, July 1995. http://ieeexplore.ieee.org/document/504983/
12. B. S. Krongold and D. L. Jones, "An active-set approach for OFDM PAR reduction via tone reservation," IEEE Trans. Signal Processing, vol.52, no.2, pp.495-509, February 2004. http://ieeexplore.ieee.org/document/1261335/
13. J. Armstrong, "Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering," IEE Electron. Lett., vol.38, no.5, pp.246-247, February 2002. http://ieeexplore.ieee.org/document/990223/
14. C.-D. Chung and S.-M. Cho, "Constant-envelope orthogonal frequency division multiplexing modulation," APCC/OECC Conference, vol.1, pp.629-632, Beijing, October 1999. http://ieeexplore.ieee.org/document/824966/
15. R. Pacheco and D. Hatzinakos, "Error rate analysis of phase-modulated OFDM (OFDM-PM) in AWGN channels," IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, pp.IV337-IV-340, Toulouse, May 2006. http://ieeexplore.ieee.org/document/1660974/
16. Steve C Thompson, Ahsen U. Ahmed, John G. Proakis and James R. Zeidler, "Constant Envelope OFDM Phase Modulation: Spectral Containment Signal Space Properties and Performance," IEEE MILCOM, pp.1129-1135, Monterey, 2004. http://ieeexplore.ieee.org/document/1495013/

17. S. C. Thompson, J. G. Proakis, and J. R. Zeidler, "Constant Envelop Binary OFDM Phase Modulation," IEEE MILCOM, pp.621-626, Boston, October 2003. http://ieeexplore.ieee.org/document/1290175/

18. J. G. Proakis and M. Salehi, Communication Systems Engineering. New Jersey: Prentice-Hall, 1994.

19. R. Zeimer and W. Tranter, Principles of Communications: Systems, Modulation, and Noise, 4th ed. New York: John Wiley, 1995.

20. J. J. Downing, Modulation Systems and Noise. Prentice-Hall, 1964.

21. J. G. Proakis, Digital Communications, 4th ed. New York: McGraw- Hill, 2001.

22. M. K. Simon and M.-S. Alouini, Digital Communication over Fading Channels. New York: John Wiley and Sons, Inc., 2000.

23. I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series, and Products, vol. 6, California, Academic Press, 2000.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 261-273 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Short CCA-Secure Attribute-Based Encryption

Hiroaki Anada[*,1], Seiko Arita[2]

[1]*Department of Information Security, University of Nagasaki, 851–2195, Japan*

[2]*Graduate School of Information Security, Institute of Information Security, 221–0835, Japan*

ARTICLEINFO

Article history:
Received: 16 November, 2017
Accepted: 16 January, 2018
Online: 31 January, 2018

Keywords:
Role-based access control
Attribute-based encryption
Direct chosen-ciphertext security
Twin Diffie-Hellman

ABSTRACT

Chosen-ciphertext attacks (CCA) are typical threat on public-key encryption schemes. We show direct chosen-ciphertext security modification in the case of attribute-based encryption (ABE), where an ABE scheme secure against chosen-plaintext attacks (CPA) is converted into an ABE scheme secure against CCA by individual techniques. Our modification works in the setting that the Diffie-Hellman tuple to be verified in decryption is in the target group of a bilinear map. The employed techniques result in expansion of the secret-key length and the decryption cost by a factor of four, while the public-key and the ciphertext lengths and the encryption cost remain almost the same.

1 Introduction

Access control is one of the fundamental processes and requirements in cybersecurity. Attribute-based encryption (ABE) invented by Sahai and Waters [1], where attributes mean authorized credentials, enables to realize access control which is functionally close to role-based access control (RBAC), but *by encryption*. In key-policy ABE (KP-ABE) introduced by the subsequent work of Goyal, Pandey, Sahai and Waters [2], a secret key is associated with an access policy over attributes, while a ciphertext is associated with a set of attributes. In a dual manner, in ciphertext-policy ABE (CP-ABE) [2, 3, 4], a ciphertext is associated with an access policy over attributes, while a secret key is associated with a set of attributes. In a KP-ABE or CP-ABE scheme, a secret key works to decrypt a ciphertext if and only if the associated set of attributes satisfies the associated access policy. The remarkable feature of ABE is attribute privacy; that is, in decryption, no information about the access policy and the identity of the secret key owner in the case of KP-ABE (or, the attributes and the identity of the secret key owner in the case of CP-ABE) leaks except the fact that the set of attributes satisfies the access policy. Since the proposals, it has been studied to attain certain properties such as indistinguishability against chosen-plaintext attacks (IND-CPA) in the standard model [4] and adaptive security against adversary's choice of a target access policy [5].

In this paper[1], we work through resolving a problem of constructing a shorter ABE scheme that attains indistinguishability against chosen-ciphertext attacks (IND-CCA) in the standard model. Here CCA means that an adversary can collects decryption results of ciphertexts of its choice through adversaries' attacking. Note that "provable security" of a cryptographic primitive is now a must requirement when we employ the primitive in a system, where it means that an appropriately defined security is polynomially reduced to the hardness of a computational problem. Moreover, the CCA security of an encryption scheme is preferable to attain because the CCA security is one of the theoretically highest securities and hence the scheme can be used widely.

To capture the idea of our approach, let us recall the case of identity-based encryption (IBE). The CHK transformation of Canetti, Halevi and Katz [7] is a generic tool for obtaining IND-CCA secure IBE scheme. It transforms any hierarchical IBE (HIBE) scheme that is selective-ID IND-CPA secure [8] into an IBE scheme that is adaptive-ID IND-CCA secure [8]. A point of the CHK transformation is that it introduces a dummy identity vk that is a verification key of a one-time signature. Then a ciphertext is attached with vk and a signature σ, which is generated each time one executes encryption. In contrast, the direct chosen-ciphertext security technique for IBE of

[*]Hiroaki Anada, 1–1–1, Manabino, Nagayo-cho, Nishisonogi-gun, Nagasaki, +81-95-813-5113 & anada@sun.ac.jp

[1]This paper is an extension of the work originally presented in SMARTCOMP 2017 [6]. The schemes of KP-ABKEM and KP-ABE have been newly proposed.

https://dx.doi.org/10.25046/aj030132

Boyen, Mei and Waters [9] is individual modification for obtaining an IND-CCA secure IBE scheme. It converts a HIBE scheme that is adaptive-ID IND-CPA secure into an IBE scheme that is adaptive-ID IND-CCA secure. Though the technique needs to treat each scheme individually, the obtained scheme attains better performance than that obtained by the generic tool (the CHK transformation). Let us transfer into the case of ABE. The transformation in [10] is a generic tool for obtaining IND-CCA secure ABE scheme. It transforms any ABE scheme (with the delegatability or the verifiability [10]) that is IND-CPA secure into an ABE scheme that is IND-CCA secure. A point of their transformation is, similar to the case of IBE, that it introduces a dummy attribute vk that is a verification key of a one-time signature. Then a ciphertext is attached with vk and a signature σ. Notice here that discussing direct chosen-ciphertext security modification for ABE (in the standard model) is a missing piece. One of the reasons seems that there is an obstacle that a Diffie Hellman tuple to be verified is in the target group of a bilinear map. In that situation, the bilinear map looks of no use.

1.1 Our Contribution

A contribution is that we fill in the missing piece; we demonstrate direct chosen-ciphertext security modification in the case of the Waters CP-ABE scheme [4] and the KP-ABE scheme of Ostrovsky, Sahai and Waters [11] To overcome the above obstacle, we employ the technique of the Twin Diffie-Hellman Trapdoor Test of Cash, Kiltz and Shoup [12]. In addition, we also utilize the algebraic trick of Boneh and Boyen [13] and Kiltz [14] to reply for adversary's decryption queries.

1.2 Related Works

Waters [4] pointed out that IND-CCA security would be attained by the CHK transformation. Gorantla, Boyd and Nieto [15] constructed a IND-CCA secure CP-ABKEM in the random oracle model. In [10] the authors proposed a generic transformation of a IND-CPA secure ABE scheme into a IND-CCA secure ABE scheme. Their transformation is considered to be an ABE-version of the CHK transformation, and it is versatile. Especially, it can be applied to non-pairing-based scheme.

The Waters CP-ABE [4] can be captured as a CP-ABKEM: the blinding factor can be considered as a random one-time key. This Waters CP-ABKEM is IND-CPA secure because the Waters CP-ABE is proved to be IND-CPA secure. For theoretical simplicity, we demonstrate an individual conversion of the Waters CP-ABKEM into a CP-ABKEM which is IND-CCA secure. Then we provide a CP-ABE scheme which is IND-CCA secure. As for KP-ABE, we demonstrate an individual conversion of KP-ABKEM of Ostrovsky, Sahai and Waters [11], which is IND-CPA secure, into a KP-ABKEM which is IND-CCA secure. Then we provide a KP-ABE scheme which is IND-CCA secure.

Finally, we note that there is a remarkable work of CP-ABE schemes and KP-ABE schemes with *constant-size* ciphertexts [16, 17]. Our direct chosen-ciphertext security modification is not constant-size ciphertexts but a different approach for easier implementation in engineering.

1.3 Organization of the Paper

In Section 2, we survey concepts, definitions and techniques needed. In Section 3, we revisit the concept, the algorithm and the security of the twin Diffie-Hellman technique. In Section 4, we construct a CCA-secure CP-ABKEM from the Waters CPA-secure CP-ABKEM [4], and provide a security proof. Also, we describe the encryption version, a CCA-secure CP-ABE. In Section 5, we construct a CCA-secure KP-ABKEM from the Ostrovsky-Sahai-Waters CPA-secure KP-ABKEM [11], and provide a security proof. Also, we describe the encryption version, a CCA-secure KP-ABE. In Section 6, we compare efficiency of our CP-ABE and KP-ABE schemes with the original schemes, and also, with the schemes obtained by applying the generic transformation [10] to the original schemes. In Section 7, we conclude our work.

2 Preliminaries

The security parameter is denoted λ. A prime of bit length λ is denoted p. A multiplicative cyclic group of order p is denoted $\mathbb{G}$. The ring of exponent domain of $\mathbb{G}$, which consists of integers from 0 to $p-1$ with modulo p operation, is denoted $\mathbb{Z}_p$.

2.1 Bilinear Map

We remark first that our description in the subsequent sections is in the setting of a symmetric bilinear map for simplicity, but we can employ an asymmetric bilinear map instead for better efficiency as is noted in Section 6. Let $\mathbb{G}$ and $\mathbb{G}_T$ be two multiplicative cyclic groups of prime order p. Let g be a generator of $\mathbb{G}$ and e be a bilinear map, $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$. The bilinear map e has the following properties:
1. Bilinearity: for all $u, v \in \mathbb{G}$ and $a, b \in \mathbb{Z}_p$, we have $e(u^a, v^b) = e(u, v)^{ab}$.
2. Non-degeneracy: $e(g, g) \neq \mathrm{id}_{\mathbb{G}_T}$ (: the identity element of the group $\mathbb{G}_T$).

Parameters of a bilinear map are generated by a probabilistic polynomial time (PPT) algorithm $\mathtt{Grp}$ on input λ: $(p, \mathbb{G}, \mathbb{G}_T, g, e) \leftarrow \mathtt{Grp}(\lambda)$.

Hereafter we assume that the group operation in $\mathbb{G}$ and $\mathbb{G}_T$ and the bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$ are computable in PT in λ.

2.2 Access Structure

Let $\mathcal{U} = \{\chi_1, \ldots, \chi_u\}$ be a set of attributes, or simply set $\mathcal{U} = \{1, \ldots, u\}$ by numbering. An *access structure*, which corresponds to an access policy, is defined as a collection $\mathbb{A}$ of non-empty subsets of $\mathcal{U}$; that is, $\mathbb{A} \subset 2^{\mathcal{U}} \setminus \{\phi\}$. An access structure $\mathbb{A}$ is called *monotone* if for any $B \in \mathbb{A}$ and $B \subset C$, $C \in \mathbb{A}$ holds. The sets in $\mathbb{A}$ are called authorized sets, and the sets not in $\mathbb{A}$ are called unauthorized sets. We will consider in this paper only monotone access structures.

2.3 Linear Secret-Sharing Scheme

We only describe a linear secret-sharing scheme (LSSS) in our context of attribute-based schemes. A secret-sharing scheme Π over the attribute universe $\mathcal{U}$ is called linear over $\mathbb{Z}_p$ if:
1. The shares for each attribute form a vector over $\mathbb{Z}_p$,
2. There exists a matrix M of size $l \times n$ called the share-generating matrix for Π and a function ρ which maps each row index i of M to an attribute in $\mathcal{U} = \{1, \ldots, u\}$: $\rho : \{1, ..., l\} \to \mathcal{U}$.

To make shares, we first choose a random vector $\vec{v} = (s, y_2, \ldots, y_n) \in \mathbb{Z}_p^n$: s is a secret to be shared. For $i = 1$ to l, we calculate each share $\lambda_i = \vec{v} \cdot M_i$, where M_i denotes the i-th row vector of M and $\cdot$ denotes the formal inner product. LSSS $\Pi = (M, \rho)$ defines an access structure $\mathbb{A}$ through ρ.

Suppose that an attribute set S satisfies $\mathbb{A}$ ($S \in \mathbb{A}$) and let $I_S = \rho^{-1}(S) \subset \{1, \ldots, l\}$. Then, let $\{\omega_i \in \mathbb{Z}_p; i \in I_S\}$ be a set of constants (*linear reconstruction constants*) such that if $\{\lambda_i \in \mathbb{Z}_p; i \in I_S\}$ are valid shares of a secret s according to M, then $\sum_{i \in I_S} \omega_i \lambda_i = s$. It is known that these constants $\{\omega_i\}_{i \in I_S}$ can be found in time polynomial in l: the row size of the share-generating matrix M. If S does not satisfy $\mathbb{A}$ ($S \notin \mathbb{A}$), then no such constants $\{\omega_i\}_{i \in I_S}$ exist.

2.4 Attribute-Based Key Encapsulation Mechanism

Ciphertext-policy attribute-based key encapsulation mechanism (CP-ABKEM). A CP-ABKEM consists of four PPT algorithms (Setup, Encap, KeyGen, Decap)[2].

Setup$(\lambda, \mathcal{U})$**.** A setup algorithm Setup takes as input the security parameter λ and the attribute universe $\mathcal{U} = \{1, \ldots, u\}$. It returns a public key PK and a master secret key MSK.

Encap$(\mathbf{PK}, \mathbb{A})$**.** An encapsulation algorithm Encap takes as input the public key PK and an access structure $\mathbb{A}$. It returns a random string κ and its encapsulation ψ. Note that $\mathbb{A}$ is contained in ψ.

KeyGen$(\mathbf{PK}, \mathbf{MSK}, S)$**.** A key generation algorithm KeyGen takes as input the public key PK, the master secret key MSK and an attribute set S. It returns a secret key SK_S corresponding to S. Note that S is contained in SK_S.

Decap$(\mathbf{PK}, \mathbf{SK}_S, \psi)$**.** A decapsulation algorithm Decap takes as input the public key PK, an encapsulation (we also call it a ciphertext according to context) ψ and a secret key SK_S. It first checks whether $S \in \mathbb{A}$, where S and $\mathbb{A}$ are contained in SK_S and ψ, respectively. If the check result is False, it puts $\hat{\kappa} = \perp$. It returns a decapsulation result $\hat{\kappa}$.

Chosen-Ciphertext Attack on CP-ABKEM. According to previous works (for example, see [15]), the chosen-ciphertext attack on a CP-ABKEM is formally defined as the indistinguishability game (IND-CCA game). In this paper, we consider the *selective game on a target access structure* (IND-sel-CCA game); that is, the adversary $\mathcal{A}$ declares a target access structure $\mathbb{A}^*$ *before* $\mathcal{A}$ receives a public key PK, which is defined as the following experiment.

$\mathbf{Experiment}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U})$
$\mathbb{A}^* \leftarrow \mathcal{A}(\lambda, \mathcal{U}), (\mathrm{PK}, \mathrm{MSK}) \leftarrow \mathbf{Setup}(\lambda, \mathcal{U})$
$\epsilon \leftarrow \mathcal{A}^{\mathbf{KeyGen}(\mathrm{PK},\mathrm{MSK},\cdot),\mathbf{Decap}(\mathrm{PK},\mathrm{SK}_\cdot,\cdot)}(\mathrm{PK})$
$(\kappa^*, \psi^*) \leftarrow \mathbf{Encap}(\mathrm{PK}, \mathbb{A}^*), \kappa \leftarrow KeySp(\lambda), b \leftarrow \{0, 1\}$
If $b = 1$ then $\tilde{\kappa} = \kappa^*$ else $\tilde{\kappa} = \kappa$
$b' \leftarrow \mathcal{A}^{\mathbf{KeyGen}(\mathrm{PK},\mathrm{MSK},\cdot),\mathbf{Decap}(\mathrm{PK},\mathrm{SK}_\cdot,\cdot)}(\tilde{\kappa}, \psi^*)$
If $b' = b$ then return Win else return Lose.

In the above experiment, two kinds of queries are issued by $\mathcal{A}$. One is key-extraction queries. Indicating an attribute set S_i, $\mathcal{A}$ queries its key-extraction oracle $\mathrm{KeyGen}(\mathrm{PK}, \mathrm{MSK}, \cdot)$ for the secret key SK_{S_i}. Here we do not require any input attribute sets S_{i_1} and S_{i_2} to be distinct. Another is decapsulation queries. Indicating a pair (S_j, ψ_j) of an attribute set and an encapsulation, $\mathcal{A}$ queries its decapsulation oracle $\mathrm{Decap}(\mathrm{PK}, \mathrm{SK}_\cdot, \cdot)$ for the decapsulation result $\hat{\kappa}_j$. Here an access structure $\mathbb{A}_j$, which is used to generate an encapsulation ψ_j, is implicitly included in ψ_j. In the case that $S \notin \mathbb{A}$, $\hat{\kappa}_j = \perp$ is replied to $\mathcal{A}$. Both kinds of queries are at most q_k and q_d times in total, respectively, which are polynomial in λ.

The access structure $\mathbb{A}^*$ declared by $\mathcal{A}$ is called a *target access structure*. Two restrictions are imposed on $\mathcal{A}$ concerning $\mathbb{A}^*$. In key-extraction queries, each attribute set S_i must satisfy $S_i \notin \mathbb{A}^*$. In decapsulation queries, each pair (S_j, ψ_j) must satisfy $S_j \notin \mathbb{A}^* \vee \psi_j \neq \psi^*$.

The *advantage* of the adversary $\mathcal{A}$ over CP-ABKEM in the IND-CCA game is defined as the following probability:

$$\mathbf{Adv}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U}) \stackrel{\text{def}}{=} \Pr[\mathbf{Experiment}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U}) \text{ returns Win}].$$

CP-ABKEM is called *selectively secure against chosen-ciphertext attacks* if, for any PPT adversary $\mathcal{A}$ and for any attribute universe $\mathcal{U}$, $\mathbf{Adv}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U})$ is negligible in λ. Here we must distinguish the two cases; the case that $\mathcal{U}$ is small (i.e. $|\mathcal{U}| = u$ is bounded by a polynomial of λ) and the case that $\mathcal{U}$ is large (i.e. u

[2] In Gorantla, Boyd and Nieto [15], they say *encapsulation-policy* attribute-based-KEM (EP-AB-KEM) instead of saying ciphertext-policy attribute-based KEM here.

is not necessarily bounded by a polynomial of λ). We assume the *small case* in this paper.

In the indistinguishability game against *chosen-plaintext attack* (IND-CPA game), the adversary $\mathcal{A}$ issues no decapsulation query (that is, $q_{\mathrm{d}} = 0$).

Ciphertext-Policy Attribute-Based Encryption Scheme (CP-ABE). In the case of the encryption version (i.e. CP-ABE), Encap(PK, $\mathbb{A}$) and Decap(PK, SK_S, ψ) are replaced by PPT algorithms Encrypt(PK, $\mathbb{A}$, m) and Decrypt(PK, SK_S, CT), respectively, where m and CT mean a message and a ciphertext, respectively.

The IND-CCA game for CP-ABE is defined in the same way as for CP-ABKEM above, except the following difference. In Challenge phase, the adversary $\mathcal{A}$ submits two equal length messages (plaintexts) m_0 and m_1. Then the challenger flips a coin $b \in \{0,1\}$ and gives an encryption result CT of m_b to $\mathcal{A}$. In Guess phase, the adversary $\mathcal{A}$ returns $b' \in \{0,1\}$. If $b' = b$, then $\mathcal{A}$ wins in the IND-CCA game. Otherwise, $\mathcal{A}$ loses.

Key-Policy Attribute-Based Key Encapsulation Mechanism (KP-ABKEM) and Encryption Scheme (KP-ABE). The *key-policy* case is analogously defined as the case of the ciphertext-policy case. We state only the syntax and the security experiment of the key-policy ABKEM.

Setup($\lambda, \mathcal{U}$)**.** A setup algorithm Setup takes as input the security parameter λ and the attribute universe $\mathcal{U} = \{1, \ldots, u\}$. It returns a public key PK and a master secret key MSK.

Encap(**PK**, S)**.** An encapsulation algorithm Encap takes as input the public key PK and an attribute set S. It returns a random string κ and its encapsulation ψ. Note that S is contained in ψ.

KeyGen(**PK**, **MSK**, $\mathbb{A}$)**.** A key generation algorithm KeyGen takes as input the public key PK, the master secret key MSK and an access structure $\mathbb{A}$. It returns a secret key $\mathrm{SK}_{\mathbb{A}}$ corresponding to S. Note that $\mathbb{A}$ is contained in $\mathrm{SK}_{\mathbb{A}}$.

Decap(**PK**, **SK**$_{\mathbb{A}}$, ψ)**.** A decapsulation algorithm Decap takes as input the public key PK, an encapsulation (we also call it a ciphertext according to context) ψ and a secret key $\mathrm{SK}_{\mathbb{A}}$. It first checks whether $S \in \mathbb{A}$. If the check result is False, it puts $\hat{\kappa} = \perp$. It returns a decapsulation result $\hat{\kappa}$.

Chosen-Ciphertext Attack on KP-ABKEM. The *selective game on a target attribute set* (IND-sel-CCA game) is defined by the following experiment.

$\mathbf{Experiment}^{\text{ind-sel-cca}}_{\mathcal{A},\mathsf{KP-ABKEM}}(\lambda, \mathcal{U})$
$S^* \leftarrow \mathcal{A}(\lambda, \mathcal{U}), (\mathrm{PK}, \mathrm{MSK}) \leftarrow \mathbf{Setup}(\lambda, \mathcal{U})$
$\epsilon \leftarrow \mathcal{A}^{\mathbf{KeyGen}(\mathrm{PK},\mathrm{MSK},\cdot),\mathbf{Decap}(\mathrm{PK},\mathrm{SK}_{\cdot},\cdot)}(\mathrm{PK})$
$(\kappa^*, \psi^*) \leftarrow \mathbf{Encap}(\mathrm{PK}, S^*), \kappa \leftarrow \mathit{KeySp}(\lambda), b \leftarrow \{0,1\}$
If $b = 1$ then $\tilde{\kappa} = \kappa^*$ else $\tilde{\kappa} = \kappa$
$b' \leftarrow \mathcal{A}^{\mathbf{KeyGen}(\mathrm{PK},\mathrm{MSK},\cdot),\mathbf{Decap}(\mathrm{PK},\mathrm{SK}_{\cdot},\cdot)}(\tilde{\kappa}, \psi^*)$
If $b' = b$ then return Win else return Lose.

2.5 Target Collision Resistant Hash Functions

Target collision resistant (TCR) hash functions [18] are treated as a family. Let us denote a function family as $\mathit{Hfam}(\lambda) = \{H_\mu\}_{\mu \in \mathit{HKey}(\lambda)}$. Here $\mathit{HKey}(\lambda)$ is a hash key space, $\mu \in \mathit{HKey}(\lambda)$ is a hash key and H_μ is a function from $\{0,1\}^*$ to $\{0,1\}^\lambda$. We may assume that H_μ is from $\{0,1\}^*$ to $\mathbb{Z}_p$, where p is a prime of length λ.

Given a PPT algorithm $\mathcal{CF}$, a collision finder, we consider the following experiment (the target collision resistance game).

$\mathbf{Experiment}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda)$
$m^* \leftarrow \mathcal{CF}(\lambda), \mu \leftarrow \mathit{HKey}(\lambda), m \leftarrow \mathcal{CF}(\mu)$
If $m^* \neq m \wedge H_\mu(m^*) = H_\mu(m)$
then return Win else return Lose.

Then we define *$\mathcal{CF}$'s advantage over Hfam in the game of target collision resistance* as follows.

$$\mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda) \stackrel{\mathrm{def}}{=} \Pr[\mathbf{Experiment}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda) \text{ returns Win}].$$

We say that *Hfam* is *a TCR function family* if, for any PPT algorithm $\mathcal{CF}$, $\mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda)$ is negligible in λ.

TCR hash function families can be constructed based on the existence of a one-way function [18].

3 The Twin Diffie-Hellman Technique Revisited

A 6-tuple $(g, X_1, X_2, Y, Z_1, Z_2) \in \mathbb{G}^6$ is called a *twin Diffie-Hellman tuple* if the tuple is written as $(g, g^{x_1}, g^{x_2}, g^y, g^{x_1 y}, g^{x_2 y})$ for some elements x_1, x_2, y in $\mathbb{Z}_p$. In other words, a 6-tuple $(g, X_1, X_2, Y, Z_1, Z_2)$ is a twin Diffie-Hellman tuple (twin DH tuple, for short) if $Y = g^y$ and $Z_1 = X_1^y$ and $Z_2 = X_2^y$.

The following lemma of Cash, Kiltz and Shoup will be used in the security proof to decide whether a tuple is a twin DH tuple or not.

Lemma 1 ("Trapdoor Test"[12]) *Let X_1, r, s be mutually independent random variables, where X_1 takes values in $\mathbb{G}$, and each of r, s is uniformly distributed over $\mathbb{Z}_p$. Define the random variable $X_2 = X_1^{-r} g^s$. Suppose that $\hat{Y}, \hat{Z}_1, \hat{Z}_2$ are random variables taking values in $\mathbb{G}$, each of which is defined independently of r. Then the probability that the truth value of $\hat{Z}_1{}^r \hat{Z}_2 = \hat{Y}^s$ does not agree with the truth value of $(g, X_1, X_2, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ being a twin DH tuple is at most $1/p$. Moreover, if $(g, X_1, X_2, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ is a twin DH tuple, then $\hat{Z}_1{}^r \hat{Z}_2 = \hat{Y}^s$ certainly holds.*

Note that Lemma 1 is a statistical property. Especially, Lemma 1 holds without any number theoretic assumption. To be precise, we consider the following experiment of an algorithm *Cheat* with *unbounded computational power (not limited to PPT)*, where *Cheat*, given a triple (g, X_1, X_2), tries to complete a 6-tuple $(g, X_1, X_2, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ which passes the "Trapdoor Test" but which is *not* a twin DH tuple.

$$\begin{aligned}&\textbf{Experiment}^{\text{twinDH-test}}_{Cheat,\mathbb{G}}(\lambda)\\&(g,X_1)\leftarrow\mathbb{G}^2,(r,s)\leftarrow\mathbb{Z}_p^2,X_2=X_1^{-r}g^s\\&\mathbb{G}^3\ni(\hat{Y},\hat{Z}_1,\hat{Z}_2)\leftarrow Cheat(g,X_1,X_2)\\&\text{If }\hat{Z}_1{}^r\hat{Z}_2=\hat{Y}^s\wedge(g,X_1,X_2,\hat{Y},\hat{Z}_1,\hat{Z}_2)\\&\qquad\text{is NOT a twin DH tuple,}\\&\qquad\text{then return Win else return Lose}\end{aligned}$$

Let us define the advantage of *Cheat* over $\mathbb{G}$ as follows.

$$\begin{aligned}&\textbf{Adv}^{\text{twinDH-test}}_{Cheat,\mathbb{G}}(\lambda)\\&\stackrel{\text{def}}{=}\Pr[\textbf{Experiment}^{\text{twinDH-test}}_{Cheat,\mathbb{G}}(\lambda)\text{ returns Win}].\end{aligned}$$

Now we are ready to complement Lemma 1.

Lemma 2 (Complement for "Trapdoor Test" [12]) *For any algorithm* Cheat *with unbounded computational power,* $\textbf{Adv}^{twinDH\text{-}test}_{\text{Cheat},\mathbb{G}}(\lambda)$ *is at most* $1/p$.

For a proof of Lemma 2, see Appendix A.

4 Securing the Waters CP-ABKEM against Chosen-Ciphertext Attacks

In this section, we describe our direct chosen-ciphertext security technique by applying it to the Waters CP-ABE [4].

Overview of Our Modification The Waters CP-ABE is proved to be secure in the IND-sel-CPA game [4]. We convert it into a scheme that is secure in the IND-sel-CCA game by employing the Twin Diffie-Hellman technique of Cash, Kiltz and Shoup [12] and the algebraic trick of Boneh and Boyen [13] and Kiltz [14].

In encryption, a ciphertext becomes to contain additional two elements (d_1,d_2), which function in decryption as a "check sum" to verify that a tuple is certainly a twin DH tuple.

In security proof, the Twin Diffie-Hellman Trapdoor Test does the function instead. It is noteworthy that we are unable to use the bilinear map instead because the tuple to be verified is in the target group. In addition, the algebraic trick enables to answer for adversary's decryption queries. Note also that the both technique become compatible by introducing random variables.

Key Encapsulation and Encryption. The Waters CP-ABE can be captured as a CP-ABKEM: the blinding factor of the form $e(g,g)^{\alpha s}$ in the Waters CP-ABE can be considered as a random one-time key. So we call it the Waters CP-ABKEM hereafter and denote it as $\texttt{CP-ABKEM}_{\text{cpa}}$. Likewise, we distinguish parameters and algorithms of $\texttt{CP-ABKEM}_{\text{cpa}}$ by the index $_{\text{cpa}}$. For theoretical simplicity, we first develop a KEM CP-ABKEM.

4.1 Our Construction

Our CP-ABKEM consists of the following four PPT algorithms (Setup, Encap, KeyGen, Decap). Roughly speaking, the Waters original scheme $\texttt{CP-ABKEM}_{\text{cpa}}$ (the first scheme in [4]) corresponds to the case $k=1$ below excluding the "check sum" (d_1,d_2).

Setup$(\lambda,\mathcal{U})$**.** Setup takes as input the security parameter λ and the attribute universe $\mathcal{U}=\{1,\ldots,u\}$. It runs $\texttt{Grp}(\lambda)$ to get $(p,\mathbb{G},\mathbb{G}_T,g,e)$, where $\mathbb{G}$ and $\mathbb{G}_T$ are cyclic groups of order p, $e:\mathbb{G}\times\mathbb{G}\to\mathbb{G}_T$ is a bilinear map and g is a generator of $\mathbb{G}$. These become public parameters. Then Setup chooses u random group elements $h_1,\ldots,h_u\in\mathbb{G}$ that are associated with the u attributes. In addition, it chooses random exponents $\alpha_k\in\mathbb{Z}_p,k=1,\ldots,4,a\in\mathbb{Z}_p$ and a hash key $\eta\in HKey(\lambda)$. The public key is published as $\text{PK}=(g,g^a,h_1,\ldots,h_u,e(g,g)^{\alpha_1},\ldots,e(g,g)^{\alpha_4},\eta)$. The authority sets $\text{MSK}=(g^{\alpha_1},\ldots,g^{\alpha_4})$ as the master secret key.

Encap$(\textbf{PK},\mathbb{A})$**.** The encapsulation algorithm Encap takes as input the public key PK and an LSSS access structure $\mathbb{A}=(M,\rho)$, where M is an $l\times n$ matrix and ρ is the function which maps each row index i of M to an attribute in $\mathcal{U}=\{1,\ldots,u\}$. Encap first chooses a random value $s\in\mathbb{Z}_p$ that is the encryption randomness, and chooses random values $y_2,\ldots,y_n\in\mathbb{Z}_p$. Then Encap forms a vector $\vec{v}=(s,y_2,\ldots,y_n)$. For $i=1$ to l, it calculates $\lambda_i=\vec{v}\cdot M_i$, where M_i denotes the i-th row vector of M. In addition, Encap chooses random values $r_1,\ldots,r_l\in\mathbb{Z}_p$. Then, a pair of a random one-time key and its encapsulation (κ,ψ) is computed as follows.

$$\begin{aligned}&\text{Put }C'=g^s;\text{For }i=1\text{ to }l:C_i=g^{a\lambda_i}h_{\rho(i)}^{-r_i},D_i=g^{r_i};\\&\psi_{\text{cpa}}=(\mathbb{A},C',((C_i,D_i);i=1,\ldots,l)),\tau\leftarrow H_\eta(\psi_{\text{cpa}});\\&\text{For }k=1\text{ to }4:\kappa_k=e(g,g)^{\alpha_k s};d_1=\kappa_1^\tau\kappa_3,d_2=\kappa_2^\tau\kappa_4;\\&(\kappa,\psi)=(\kappa_1,(\psi_{\text{cpa}},d_1,d_2)).\end{aligned}$$

KeyGen$(\textbf{MSK},\textbf{PK},S)$**.** The key generation algorithm KeyGen takes as input the master secret key MSK, the public key PK and a set S of attributes. KeyGen first chooses a random $t_k\in\mathbb{Z}_p,k=1,\ldots,4$. It generates the secret key SK_S as follows.

$$\begin{aligned}&\text{For }k=1\text{ to }4:K_k=g^{\alpha_k}g^{at_k},L_k=g^{t_k}\\&\quad\text{For }x\in S:K_{k,x}=h_x^{t_k};\\&\text{SK}_S=((K_k,L_k,(K_{k,x};x\in S));k=1,\ldots,4).\end{aligned}$$

Decap$(\textbf{PK},\psi,\textbf{SK}_S)$**.** The decapsulation algorithm Decap takes as input the public key PK, an encapsulation ψ for an access structure $\mathbb{A}=(M,\rho)$ and a private key SK_S for an attribute set S. It first checks whether $S\in\mathbb{A}$. If the result is False, put $\hat{\kappa}=\perp$. Otherwise, let $I_S=\rho^{-1}(S)\subset\{1,\ldots,l\}$ and let $\{\omega_i\in\mathbb{Z}_p;i\in I_S\}$ be a set of linear reconstruction constants. Then, the decapsu-

lation $\hat{\kappa}$ is computed as follows.

Parse ψ into $(\psi_{\mathrm{cpa}} = (\mathbb{A}, C', ((C_i, D_i); i = 1, \ldots, l)), d_1, d_2)$;
$\tau \leftarrow H_\eta(\psi_{\mathrm{cpa}})$;
For $k = 1$ to 4 :

$$\hat{\kappa_k} = e(C', K_k) / \prod_{i \in I_S} (e(L_k, C_i) e(D_i, K_{k,\rho(i)}))^{\omega_i} = e(g,g)^{\alpha_k s}$$

If ${\hat{\kappa_1}}^{\tau} \hat{\kappa_3} \neq d_1 \vee {\hat{\kappa_2}}^{\tau} \hat{\kappa_4} \neq d_2$,
then put $\hat{\kappa} = \perp$, else put $\hat{\kappa} = \hat{\kappa_1}$.

4.2 Security and Proof

Theorem 1 *If the Waters CP-ABKEM$_{cpa}$ [4] is selectively secure against chosen-plaintext attacks and an employed hash function family Hfam has target collision resistance, then our CP-ABKEM is selectively secure against chosen-ciphertext attacks. More precisely, for any given PPT adversary $\mathcal{A}$ that attacks CP-ABKEM in the IND-sel-CCA game where decapsulation queries are at most q_d times, and for any small attribute universe $\mathcal{U}$, there exist a PPT adversary $\mathcal{B}$ that attacks CP-ABKEM$_{cpa}$ in the IND-sel-CPA game and a PPT target collision finder $\mathcal{CF}$ on Hfam that satisfy the following tight reduction.*

$$\begin{aligned} &\mathbf{Adv}^{ind\text{-}sel\text{-}cca}_{\mathcal{A},CP\text{-}ABKEM}(\lambda, \mathcal{U}) \\ \leq &\mathbf{Adv}^{ind\text{-}sel\text{-}cpa}_{\mathcal{B},CP\text{-}ABKEM_{cpa}}(\lambda, \mathcal{U}) + \mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda) + \frac{q_d}{p}. \end{aligned}$$

Proof. Given any adversary $\mathcal{A}$ that attacks our scheme CP-ABKEM in the IND-sel-CCA game, we construct an adversary $\mathcal{B}$ that attacks the Waters scheme CP-ABKEM$_{\mathrm{cpa}}$ in the IND-sel-CPA game as follows.

Commit to a Target Access Structure. $\mathcal{B}$ is given $(\lambda, \mathcal{U})$ as inputs, where λ is the security parameter and $\mathcal{U} = \{1, \ldots, u\}$ is the attribute universe. $\mathcal{B}$ invokes $\mathcal{A}$ on input $(\lambda, \mathcal{U})$ and gets a target access structure $\mathbb{A}^* = (M^*, \rho^*)$ from $\mathcal{A}$, where M^* is of size $l^* \times n^*$. $\mathcal{B}$ uses $\mathbb{A}^*$ as the target access structure of itself and outputs $\mathbb{A}^*$.

Set up. In return to outputting $\mathbb{A}^*$, $\mathcal{B}$ receives the public key $\mathrm{PK}_{\mathrm{cpa}}$ for CP-ABKEM$_{\mathrm{cpa}}$, which consists of the following components.

$$\mathrm{PK}_{\mathrm{cpa}} = (g, g^a, h_1, \ldots, h_u, e(g,g)^\alpha).$$

To set up a public key PK for CP-ABKEM, $\mathcal{B}$ *herein* needs a challenge instance: $\mathcal{B}$ queries its challenger and gets a challenge instance $(\tilde{\kappa}, \psi^*_{\mathrm{cpa}})$. It consists of the following components.

$\tilde{\kappa} = e(g,g)^{\alpha s^*}$ OR a random one-time key $\kappa \in \mathit{KeySp}(\lambda)$,
$\psi^*_{\mathrm{cpa}} = (\mathbb{A}^*, C'^* = g^{s^*}, ((C_i^*, D_i^*); i = 1, \ldots, l^*))$.

Then $\mathcal{B}$ makes the rest of parameters of PK as follows.

Choose $\eta \leftarrow \mathit{HKey}(\lambda)$ and take $\tau^* \leftarrow H_\eta(\psi^*_{\mathrm{cpa}})$;
Put $e(g,g)^{\alpha_1} = e(g,g)^\alpha$;
Choose $\gamma_1, \gamma_2 \leftarrow \mathbb{Z}_p$, put $e(g,g)^{\alpha_2} = e(g,g)^{\gamma_2} / e(g,g)^{\alpha_1 \gamma_1}$;
Choose $\mu_1, \mu_2 \leftarrow \mathbb{Z}_p$, put $e(g,g)^{\alpha_3} = e(g,g)^{\mu_1} / e(g,g)^{\alpha_1 \tau^*}$,

$$e(g,g)^{\alpha_4} = e(g,g)^{\mu_2} / e(g,g)^{\alpha_2 \tau^*}.$$

Note we have implicitly set relations in the exponent domain:

$$\begin{aligned} &\alpha_2 = \gamma_2 - \alpha_1 \gamma_1, \quad \alpha_3 = \mu_1 - \alpha_1 \tau^*, \\ &\alpha_4 = \mu_2 - \alpha_2 \tau^* = \mu_2 - (\gamma_2 - \alpha_1 \gamma_1) \tau^*. \end{aligned} \tag{1}$$

A public key PK for CP-ABKEM become:

$$\mathrm{PK} = (\mathrm{PK}_{\mathrm{cpa}}, e(g,g)^{\alpha_2}, e(g,g)^{\alpha_3}, e(g,g)^{\alpha_4}, \eta).$$

Then $\mathcal{B}$ inputs PK into $\mathcal{A}$. Note that PK determines the corresponding MSK uniquely.

Phase 1. $\mathcal{B}$ answers for two types of $\mathcal{A}$'s queries as follows.

(1) Key-Extraction Queries. In the case that $\mathcal{A}$ issues a key-extraction query for an attribute set $S \subset \mathcal{U}$, $\mathcal{B}$ has to simulate $\mathcal{A}$'s challenger. To do so, $\mathcal{B}$ issues key-extraction queries to $\mathcal{B}$'s challenger for S *repeatedly up to four times*. As replies, $\mathcal{B}$ gets four secret keys of the Waters CP-ABKEM$_{\mathrm{cpa}}$ for a single attribute set S:

$$\mathrm{SK}_{\mathrm{cpa},S,k} = (K_{\mathrm{cpa},k}, L_{\mathrm{cpa},k}, (K_{\mathrm{cpa},k,x}; x \in \mathcal{S})), k = 1, \ldots, 4.$$

We remark that, according to the randomness in the key-generation algorithm of the Waters CP-ABKEM$_{\mathrm{cpa}}$, all four secret keys $\mathrm{SK}_{\mathrm{cpa},S,1}, \ldots, \mathrm{SK}_{\mathrm{cpa},S,4}$ are random and mutually independent. To reply a secret key SK_S of our CP-ABKEM to $\mathcal{A}$, $\mathcal{B}$ converts the four secret keys as follows.

$$\begin{aligned} &K_1 = K_{\mathrm{cpa},1}, && L_1 = L_{\mathrm{cpa},1},\ K_{1,x} = K_{\mathrm{cpa},1,x}, x \in S; \\ &K_2 = g^{\gamma_2} K_{\mathrm{cpa},2}^{-\gamma_1}, && L_2 = L_{\mathrm{cpa},2}^{-\gamma_1},\ K_{2,x} = K_{\mathrm{cpa},2,x}^{-\gamma_1},\ x \in S; \\ &K_3 = g^{\mu_1} K_{\mathrm{cpa},3}^{-\tau^*}, && L_3 = L_{\mathrm{cpa},3}^{-\tau^*},\ K_{3,x} = K_{\mathrm{cpa},3,x}^{\tau^*},\ x \in S; \\ &K_4 = g^{\mu_2 - \gamma_2 \tau^*} K_{\mathrm{cpa},4}^{\gamma_1 \tau^*}, && L_4 = L_{\mathrm{cpa},4}^{\gamma_1 \tau^*},\ K_{4,x} = K_{\mathrm{cpa},4,x}^{\gamma_1 \tau^*},\ x \in S. \end{aligned}$$

Then $\mathcal{B}$ replies $\mathrm{SK}_S = ((K_k, L_k, (K_{k,x}; x \in S)); k = 1, \ldots, 4)$ to $\mathcal{A}$.

(2) Decapsulation Queries. In the case that $\mathcal{A}$ issues a decapsulation query for (S, ψ), where $S \subset \mathcal{U}$ is an attribute set and $\psi = (\psi_{\mathrm{cpa}}, d_1, d_2)$ is an encapsulation concerning $\mathbb{A}$, $\mathcal{B}$ has to simulate $\mathcal{A}$'s challenger. To do so, $\mathcal{B}$ computes the decapsulation result $\hat{\kappa}$ as follows.

If $S \notin \mathbb{A}$ then put $\hat{\kappa} = \perp$,
else
  $\tau \leftarrow H_\eta(\psi_{\mathrm{cpa}})$;
  $\hat{Y} = e(C', g)^{\tau - \tau^*}, \hat{Z}_1 = d_1 / e(C', g)^{\mu_1}, \hat{Z}_2 = d_2 / e(C', g)^{\mu_2}$;
  If ${\hat{Z}_1}^{\gamma_1} \hat{Z}_2 \neq \hat{Y}^{\gamma_2}$ (: call this checking TwinDH-Test)
    then put $\hat{\kappa} = \hat{\kappa_1} = \perp$
  else
    If $\tau = \tau^*$ then abort (: call this case Abort)
    else $\hat{\kappa} = \hat{\kappa_1} = {\hat{Z}_1}^{1/(\tau - \tau^*)}$.

Challenge. In the case that $\mathcal{A}$ queries its challenger for a challenge instance, $\mathcal{B}$ makes a challenge instance as follows.

$$\begin{aligned} &\text{Put } d_1^* = e(C'^*, g)^{\mu_1}, d_2^* = e(C'^*, g)^{\mu_2}; \\ &\text{Put } \psi^* = (\psi^*_{\mathrm{cpa}}, d_1^*, d_2^*). \end{aligned}$$

Then $\mathcal{B}$ feeds $(\tilde{\kappa}, \psi^*)$ to $\mathcal{A}$ as a challenge instance.
Phase 2. The same as in Phase 1.
Guess. In the case that $\mathcal{A}$ returns $\mathcal{A}$'s guess $\tilde{b}$, $\mathcal{B}$ returns $\tilde{b}$ itself as $\mathcal{B}$'s guess.

In the above construction of $\mathcal{B}$, $\mathcal{B}$ can perfectly simulate the real view of $\mathcal{A}$ until the case Abort happens, except for a negligible case, and hence the algorithm $\mathcal{A}$ works as designed. To see the perfect simulation with a negligible exceptional case, we are enough to prove the following seven claims.

Claim 1 *The reply $SK_S = ((K_k, L_k, (K_{k,x}; x \in S)); k = 1, \ldots, 4)$ for a key-extraction query of $\mathcal{A}$ is a perfect simulation.*

Proof. We must consider the implicit relations (1). For the index 2, we have implicitly set the randomness $t_2 = t_{\mathrm{cpa},2}(-\gamma_1)$ and we get:

$$\begin{aligned}
K_2 &= g^{\gamma_2} K_{\mathrm{cpa},2}^{-\gamma_1} = g^{\gamma_2} (g^{\alpha_1} g^{a t_{\mathrm{cpa},2}})^{-\gamma_1} \\
&= g^{\gamma_2} (g^{\alpha_1} g^{a t_2/(-\gamma_1)})^{-\gamma_1} = g^{\gamma_2 - \alpha_1 \gamma_1} g^{a t_2} = g^{\alpha_2} g^{a t_2}, \\
L_2 &= L_{\mathrm{cpa},2}^{-\gamma_1} = (g^{t_{\mathrm{cpa},2}})^{-\gamma_1} = g^{t_2}, \\
K_{2,x} &= K_{\mathrm{cpa},2,x}^{-\gamma_1} = (h_x^{t_{\mathrm{cpa},2}})^{-\gamma_1} = h_x^{t_2}, x \in S.
\end{aligned}$$

For the index 3 and 4, see Appendix B.

Claim 2 *$(e(g,g),\ e(g,g)^{\alpha_1},\ e(g,g)^{\alpha_2},\ \hat{Y},\ \hat{Z}_1,\ \hat{Z}_2)$ is a twin Diffie-Hellman tuple if and only if $(e(g,g),\ e(g,g)^{\alpha_1 \tau} e(g,g)^{\alpha_3},\ e(g,g)^{\alpha_2 \tau} e(g,g)^{\alpha_4},\ e(C', g),\ d_1,\ d_2)$ is a twin Diffie-Hellman tuple.*

Proof. This claim can be proved by a short calculation. See Appendix C. □

Claim 3 *If $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ is a twin Diffie-Hellman tuple, then $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$ certainly passes the* TwinDH-Test*: $\hat{Z}_1^{\gamma_1} \hat{Z}_2 = \hat{Y}^{\gamma_2}$.*

Proof. This claim is a direct consequence of Lemma 1. □

Claim 4 *Consider the following event which we name as* Overlook$_i$*:*

In the i-th TwinDH-Test*, the following condition holds:*

$$\begin{cases} \hat{Z}_1^{\gamma_1} \hat{Z}_2 = \hat{Y}^{\gamma_2} \text{ holds and} \\ (e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2) \\ \quad \text{is NOT a twin DH tuple.} \end{cases}$$

Then, for at most q_d times decapsulation queries of $\mathcal{A}$, the probability that at least one Overlook$_i$ *occurs is negligible in λ. More precisely, the following inequality holds:*

$$\Pr[\bigvee_{i=1}^{q_d} \textsc{Overlook}_i] \leq q_d/p. \tag{2}$$

Proof. To apply Lemma 2, we construct an algorithm $Cheat_{\lambda,\mathcal{U}}$ with unbounded computational power, which takes as input $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2})$ and returns $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$ employing the adversary $\mathcal{A}$ as a subroutine. Fig. 1 shows the construction.

First, note that the view of $\mathcal{A}$ in $Cheat_{\lambda,\mathcal{U}}$ is the same as the real view of $\mathcal{A}$ and hence the algorithm $\mathcal{A}$ works as designed.

Second, note that the return $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$ of $Cheat_{\lambda,\mathcal{U}}$ is randomized in TABLE. Hence:

$$\begin{aligned}
&\sum_{i=1}^{q_\mathrm{d}} \frac{1}{q_\mathrm{d}} \Pr[\textsc{Overlook}_i] = \frac{1}{q_\mathrm{d}} \sum_{i=1}^{q_\mathrm{d}} \Pr[\textsc{Overlook}_i] \\
&= \mathbf{Adv}^{\text{twinDH-test}}_{Cheat_{\lambda,\mathcal{U}},\mathbb{G}}(\lambda).
\end{aligned} \tag{3}$$

Third, applying Lemma 2 to $Cheat_{\lambda,\mathcal{U}}$, we get:

$$\mathbf{Adv}^{\text{twinDH-test}}_{Cheat_{\lambda,\mathcal{U}},\mathbb{G}}(\lambda) \leq 1/p. \tag{4}$$

Combining (3) and (4), we have:

$$\begin{aligned}
&\Pr[\bigvee_{i=1}^{q_\mathrm{d}} \textsc{Overlook}_i] \leq \sum_{i=1}^{q_\mathrm{d}} \Pr[\textsc{Overlook}_i] \\
&\leq q_\mathrm{d} \mathbf{Adv}^{\text{twinDH-test}}_{Cheat_{\lambda,\mathcal{U}},\mathbb{G}}(\lambda) \leq \frac{q_\mathrm{d}}{p}.
\end{aligned}$$

Claim 5 *The probability that* Overlook$_i$ *never occurs in* TwinDH-Test *for every i and* Abort *occurs is negligible in λ. More precisely, the following inequality holds:* □

$$\Pr[\Big(\bigwedge_{i=1}^{q_d} \neg \textsc{Overlook}_i\Big) \wedge \textsc{Abort}] \leq \mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda). \tag{5}$$

Proof. This claim is proved by constructing a collision finder $\mathcal{CF}$ on *Hfam*. See Appendix D. □

Claim 6 *The reply $\hat{\kappa}$ to $\mathcal{A}$ as an answer for a decapsulation query is correct.*

Claim 7 *The challenge instance $\psi^* = (\psi^*_{cpa}, d_1^*, d_2^*)$ is correctly distributed.*

Proof. These two claims are proved by a direct calculation. See Appendices E and F, respectively. □

Evaluation of the Advantage of $\mathcal{B}$. Now we are ready to evaluate the advantage of $\mathcal{B}$ in the IND-sel-CPA game. That $\mathcal{A}$ wins in the IND-sel-CCA game means that $(\tilde{\kappa}, \psi^* = (\psi^*_{\mathrm{cpa}}, d_1^*, d_2^*))$ is correctly guessed. This is equivalent to that $(\tilde{\kappa}, \psi^*_{\mathrm{cpa}})$ is correctly guessed because ψ^*_{cpa} determines the consistent blinding factor $\kappa^* = e(g,g)^{\alpha s^*}$ uniquely. This means that $\mathcal{B}$ wins in the IND-sel-CPA game.

Therefore, the probability that $\mathcal{B}$ wins is equal to the probability that $\mathcal{A}$ wins, Overlook$_i$ never holds in TwinDH-Testfor each i and Abort never occurs. So

Given $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2})$ as input :
Set up
Initialize the inner state, put TABLE = ϕ;
Get a target access structure $\mathbb{A}^* \leftarrow \mathcal{A}(\lambda, \mathcal{U})$;
Compute the base $g \in \mathbb{G}$ from $(e(g,g), e)$;
Choose $a \in \mathbb{Z}_p$ and $h_1, \ldots, h_u \in \mathbb{G}$;
Put $\mathrm{PK}_{\mathrm{cpa}} = (g, g^a, h_1, \ldots, h_u, e(g,g)^{\alpha_1})$;
Get $(\kappa^*, \psi^*_{\mathrm{cpa}}) \leftarrow \mathbf{Encap}_{\mathrm{cpa}}(\mathrm{PK}_{\mathrm{cpa}}, \mathbb{A}^*)$;
Choose $\eta \leftarrow \mathit{HKey}(\lambda)$ and compute $\tau^* \leftarrow H_\eta(\psi^*_{\mathrm{cpa}})$;
Compute discrete logarithms $\alpha_1, \alpha_2 \in \mathbb{Z}_p$ of $e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}$ to the base $e(g,g)$;
Choose $\mu_1, \mu_2 \leftarrow \mathbb{Z}_p$, put $\alpha_3 = \mu_1 - \alpha_1\tau^*, \alpha_4 = \mu_2 - \alpha_2\tau^*$;
Put $\mathrm{PK} = (\mathrm{PK}_{\mathrm{cpa}}, e(g,g)^{\alpha_2}, e(g,g)^{\alpha_3}, e(g,g)^{\alpha_4}, \eta), \mathrm{MSK} = (g^{\alpha_1}, g^{\alpha_2}, g^{\alpha_3}, g^{\alpha_4})$;
Give PK to $\mathcal{A}$;
Phase 1
In the case that $\mathcal{A}$ makes a key-extraction query for $S \subset \mathcal{U}$;
Reply SK_S to $\mathcal{A}$ in the same way as **KeyGen** does using MSK;
In the case that $\mathcal{A}$ makes a decapsulation query for $(\mathbb{A}, \psi = (\psi_{\mathrm{cpa}}, d_1, d_2), S)$;
Reply $\hat{\kappa}$ to $\mathcal{A}$ in the same way as **Decap** does using MSK;
Compute $\hat{Y} = e(C', g)^{\tau - \tau^*}, \hat{Z}_1 = d_1 / e(C', g)^{\mu_1}, \hat{Z}_2 = d_2 / e(C', g)^{\mu_2}$;
Update TABLE = TABLE $\cup (\hat{Y}, \hat{Z}_1, \hat{Z}_2)$;
Challenge
In the case that $\mathcal{A}$ makes a challenge instance query;
Put $d^*_1 = e(C'^*, g)^{\mu_1}, d^*_2 = e(C'^*, g)^{\mu_2}, \psi^* = (\psi^*_{\mathrm{cpa}}, d^*_1, d^*_2)$;
Choose $\kappa \leftarrow \mathit{KeySp}(\lambda), b \leftarrow \{0,1\}$;
If $b = 1$ then put $\tilde{\kappa} = \kappa^*$ else put $\tilde{\kappa} = \kappa$;
Reply $(\tilde{\kappa}, \psi^*)$ to $\mathcal{A}$;
Phase 2
The same as in Phase 1;
Return
In the case that $\mathcal{A}$ returns its guess b^*;
Choose one triple $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$ from TABLE at random;
Return $(\hat{Y}, \hat{Z}_1, \hat{Z}_2)$.

Figure 1: An Algorithm $\mathit{Cheat}_{\lambda,\mathcal{U}}$ with Unbounded Computational Power for a Proof of Claim 4.

we have:

$$
\begin{aligned}
&\Pr[\mathcal{B} \text{ wins}] \\
&= \Pr[(\mathcal{A} \text{ wins}) \wedge \Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg\textsc{Overlook}_i\Big) \wedge (\neg\textsc{Abort})] \\
&= \Pr[\mathcal{A} \text{ wins}] \\
&\quad - \Pr[(\mathcal{A} \text{ wins}) \wedge \neg\Big(\Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg\textsc{Overlook}_i\Big) \wedge (\neg\textsc{Abort})\Big)] \\
&\geq \Pr[\mathcal{A} \text{ wins}] - \Pr[\neg\Big(\Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg\textsc{Overlook}_i\Big) \wedge (\neg\textsc{Abort})\Big)] \\
&= \Pr[\mathcal{A} \text{ wins}] \\
&\quad - (\Pr[\bigvee_{i=1}^{q_{\mathrm{d}}} \textsc{Overlook}_i] + \Pr[\Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg\textsc{Overlook}_i\Big) \wedge \textsc{Abort}]).
\end{aligned}
$$

Substituting (2), (5) and advantages into the above, we have:

$$
\begin{aligned}
&\mathbf{Adv}^{\text{ind-sel-cpa}}_{\mathcal{B},\mathtt{CP\text{-}ABKEM}_{\mathrm{cpa}}}(\lambda, \mathcal{U}) \\
&\geq \mathbf{Adv}^{\text{ind-sel-cca}}_{\mathcal{A},\mathtt{CP\text{-}ABKEM}}(\lambda, \mathcal{U}) - \frac{q_{\mathrm{d}}}{p} - \mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF},\mathit{Hfam}}(\lambda).
\end{aligned}
$$

This is what we should prove in Theorem 1. □

4.3 Encryption Scheme from KEM

It is straightforward to construct our encryption scheme CP-ABE from CP-ABKEM. The IND-sel-CCA security of CP-ABE is proved based on IND-sel-CPA security of the Waters *KEM* $\mathtt{CP\text{-}ABKEM}_{\mathrm{cpa}}$.

Setup$(\lambda, \mathcal{U})$. The same as Setup of CP-ABKEM.

Encrypt$(\mathrm{PK}, \mathbb{A}, m)$. The same as Encap of CP-ABKEM except that Encrypt multiplies m by the blinding factor κ in the group $\mathbb{G}_T$. Encrypt returns $\mathrm{CT} = (C = m\kappa, \psi = (\psi_{\mathrm{cpa}}, d_1, d_2))$.

KeyGen$(\mathrm{MSK}, \mathrm{PK}, S)$. The same as KeyGen of CP-ABKEM.

Decrypt$(\mathrm{PK}, \mathrm{CT}, \mathrm{SK}_S)$. The same as Decap of CP-ABKEM except that Decrypt divides out C by the decapsulated blinding factor $\hat{\kappa}$. Decrypt returns the result $\hat{m}$.

4.4 Security and Proof

Theorem 2 *If the Waters* $\mathtt{CP\text{-}ABKEM}_{cpa}$ *[4] is selectively secure against chosen-plaintext attacks and an employed hash function family* Hfam *has target collision resistance, then our* CP-ABE *is selectively secure against chosen-ciphertext attacks. More precisely, for any given PPT ad-*

versary $\mathcal{A}$ that attacks CP-ABE in the IND-sel-CCA game where decryption queries are at most q_d times, and for any small attribute universe $\mathcal{U}$, there exist a PPT adversary $\mathcal{B}$ that attacks CP-ABKEM$_{cpa}$ in the IND-sel-CPA game and a PPT target collision finder $\mathcal{CF}$ on Hfam that satisfy the following inequality.

$$\begin{aligned}&\mathbf{Adv}^{ind\text{-}sel\text{-}cca}_{\mathcal{A},CP\text{-}ABE}(\lambda,\mathcal{U})\\ \leq& 2\Big(\mathbf{Adv}^{ind\text{-}sel\text{-}cpa}_{\mathcal{B},CP\text{-}ABKEM_{cpa}}(\lambda,\mathcal{U})+\mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda)+\frac{q_d}{p}\Big).\end{aligned}$$

Proof. Given any adversary $\mathcal{A}$ that attacks our scheme CP-ABE in the IND-sel-CCA game, we construct an adversary $\mathcal{B}$ that attacks the Waters KEM CP-ABKEM$_{\mathrm{cpa}}$ in the IND-sel-CPA game as follows.

Commit to a Target Access Structure. The same as that of CP-ABKEM.

Set up. In return to outputting $\mathbb{A}^*$, $\mathcal{B}$ receives the public key PK$_{\mathrm{cpa}}$ for CP-ABKEM$_{\mathrm{cpa}}$. To set up a public key PK for CP-ABE, $\mathcal{B}$ *herein* needs a challenge instance: $\mathcal{B}$ queries its challenger and gets a challenge instance $(\tilde{\kappa},\psi^*_{\mathrm{cpa}})$. The rest of procedure is the same as that of CP-ABKEM, and $\mathcal{B}$ inputs PK into $\mathcal{A}$.

Phase 1. The same as that of CP-ABKEM except that $\mathcal{B}$ replies a decrypted message $\hat{m}$ to $\mathcal{A}$ for a decryption query.

Challenge. In the case that $\mathcal{A}$ submits two plaintexts (m^*_0,m^*_1) of equal length, $\mathcal{B}$ makes a challenge ciphertext CT* as follows and feeds CT* to $\mathcal{A}$.

$$\begin{aligned}&\text{Choose } b'\leftarrow\{0,1\},\text{ put } \tilde{C}^*=m^*_{b'},\tilde{\kappa};\\ &\text{Put } d^*_1=e(C'^*,g)^{\mu_1},d^*_2=e(C'^*,g)^{\mu_2};\\ &\text{Put CT}^*=(\tilde{C}^*,\psi^*=(\psi^*_{\mathrm{cpa}},d^*_1,d^*_2)).\end{aligned}$$

Phase 2. The same as in Phase 1.

Guess. In the case that $\mathcal{A}$ returns $\mathcal{A}$'s guess $\tilde{b}$, $\mathcal{B}$ returns $\tilde{b}$ as $\mathcal{B}$'s guess.

Evaluation of the Advantage of $\mathcal{B}$. A standard argument deduces a loss of tightness by a factor of 1/2. That is;

$$\begin{aligned}&\mathbf{Adv}^{\text{ind-sel-cpa}}_{\mathcal{B},\text{CP-ABKEM}_{\text{cpa}}}(\lambda,\mathcal{U})\\ \geq&\frac{1}{2}\mathbf{Adv}^{\text{ind-sel-cca}}_{\mathcal{A},\text{CP-ABE}}(\lambda,\mathcal{U})-\frac{q_{\mathrm{d}}}{p}-\mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF},Hfam}(\lambda).\square\end{aligned}$$

5 Securing the Ostrovsky-Sahai-Waters KP-ABKEM against Chosen-Ciphertext Attacks

In this section, we describe our direct chosen-ciphertext security modification by applying it to the Ostrovsky-Sahai-Waters KP-ABE [11].

Overview of Our Modification The Ostrovsky-Sahai-Waters KP-ABE is proved to be secure in the IND-sel-CPA game [11]. We convert it into a scheme that is secure in the IND-sel-CCA game by employing the Twin Diffie-Hellman technique of Cash, Kiltz and Shoup [12] and the algebraic trick of Boneh and Boyen [13] and Kiltz [14].

In encryption, a ciphertext becomes to contain additional two elements (d_1,d_2), which function in decryption as a "check sum" to verify that a tuple is certainly a twin DH tuple.

In security proof, the Twin Diffie-Hellman Trapdoor Test does the function instead. It is noteworthy that we are unable to use the bilinear map instead because the tuple to be verified is in the target group. In addition, the algebraic trick enables to answer for adversary's decryption queries. Note also that the both technique become compatible by introducing random variables.

Key Encapsulation and Encryption. The Ostrovsky-Sahai-Waters KP-ABE can be captured as a KP-ABKEM: the blinding factor of the form $e(g,g)^{a\alpha s}$ in the Ostrovsky-Sahai-Waters KP-ABE can be considered as a random one-time key. So we call it the Ostrovsky-Sahai-Waters KP-ABKEM hereafter and denote it as KP-ABKEM$_{\mathrm{cpa}}$. Likewise, we distinguish parameters and algorithms of KP-ABKEM$_{\mathrm{cpa}}$ by the index $_{\mathrm{cpa}}$. For theoretical simplicity, we first develop a KEM KP-ABKEM.

5.1 Our Construction

Our KP-ABKEM consists of the following four PPT algorithms (Setup, Encap, KeyGen, Decap). Roughly speaking, the Ostrovsky-Sahai-Waters original scheme KP-ABKEM$_{\mathrm{cpa}}$ (the first scheme in [11]) corresponds to the case $k=1$ below excluding the "check sum" (d_1,d_2).

Setup$(\lambda,\mathcal{U})$**.** Setup takes as input the security parameter λ and the attribute universe $\mathcal{U}=\{1,\ldots,u\}$. It runs $\mathsf{Grp}(\lambda)$ to get $(p,\mathbb{G},\mathbb{G}_T,g,e)$, where $\mathbb{G}$ and $\mathbb{G}_T$ are cyclic groups of order p, $e:\mathbb{G}\to\mathbb{G}_T$ is a bilinear map and g is a generator of $\mathbb{G}$. These become public parameters. Then Setup chooses u random group elements $h_1,\ldots,h_u\in\mathbb{G}$ that are associated with the u attributes. In addition, it chooses random exponents $\alpha_k\in\mathbb{Z}_p,k=1,\ldots,4,a\in\mathbb{Z}_p$ and a hash key $\eta\in HKey(\lambda)$. The public key is published as PK $=(g,g^a,h_1,\ldots,h_u,e(g,g)^{a\alpha_1},\ldots,e(g,g)^{a\alpha_4},\eta)$. The authority sets MSK $=(\alpha_1,\ldots,\alpha_4)$ as the master secret key.

Encap$(\mathbf{PK},S)$**.** The encapsulation algorithm Encap takes as input the public key PK and a set S of attributes. Encap first chooses a random value $s\in\mathbb{Z}_p$ that is the encryption randomness. Then, a pair of a random one-time key and its encapsulation (κ,ψ) is computed as follows.

$$\begin{aligned}&\text{Put } C'=g^s;\text{For } x\in S: C_x=h^s_x\\ &\psi_{\mathrm{cpa}}=(S,C',(C_x;x\in S)),\tau\leftarrow H_\eta(\psi_{\mathrm{cpa}});\\ &\text{For } k=1\text{ to }4:\kappa_k=e(g,g)^{a\alpha_k s};d_1=\kappa_1^\tau\kappa_3,d_2=\kappa_2^\tau\kappa_4;\\ &(\kappa,\psi)=(\kappa_1,(\psi_{\mathrm{cpa}},d_1,d_2)).\end{aligned}$$

KeyGen$(\mathbf{MSK},\mathbf{PK},\mathbb{A})$**.** The key generation algorithm KeyGen takes as input the master secret key MSK, the public key PK and an LSSS access structure $\mathbb{A}=(M,\rho)$, where M is an $l\times n$ matrix and ρ is the function which maps each row index i of M to an attribute in $\mathcal{U}=\{1,\ldots,u\}$. For $k=1$ to 4, KeyGen first chooses

random values $y_{k,2},\ldots,y_{k,n} \in \mathbb{Z}_p$ and forms a vector $\vec{v_k} = (\alpha_k, y_{k,2},\ldots,y_{k,n})$. Then, for $i = 1$ to l, it calculates $\lambda_{k,i} = \vec{v_k} \cdot M_i$, where M_i denotes the i-th row vector of M, and it chooses random values $r_{k,i} \in \mathbb{Z}_p$. KeyGen generates the secret key $\mathrm{SK}_{\mathbb{A}}$ as follows.

$$\begin{aligned}&\text{For } k = 1 \text{ to } 4 : \text{For } l = 1 \text{ to } l :\\ &\quad K_{k,i} = g^{a\lambda_{k,i}} h_{\rho(i)}^{r_{k,i}}, L_{k,i} = g^{r_{k,i}}\\ &\mathrm{SK}_{\mathbb{A}} = (((K_{k,i}, L_{k,i}); i = 1,\ldots,l)k = 1,\ldots,4).\end{aligned}$$

Decap$(\mathbf{PK}, \psi, \mathbf{SK}_{\mathbb{A}})$**.** The decapsulation algorithm Decap takes as input the public key PK, an encapsulation ψ for an attribute set S and a private key $\mathrm{SK}_{\mathbb{A}}$ for an access structure $\mathbb{A} = (M, \rho)$. It first checks whether $S \in \mathbb{A}$. If the result is False, put $\hat{\kappa} = \perp$. Otherwise, let $I_S = \rho^{-1}(S) \subset \{1,\ldots,l\}$ and let $\{\omega_i \in \mathbb{Z}_p; i \in I_S\}$ be a set of linear reconstruction constants. Then, the decapsulation $\hat{\kappa}$ is computed as follows.

$$\begin{aligned}&\text{Parse } \psi \text{ into } (\psi_{\mathrm{cpa}} = (S, C', (C_x; x \in S)), d_1, d_2);\\ &\tau \leftarrow H_\eta(\psi_{\mathrm{cpa}});\\ &\text{For } k = 1 \text{ to } 4 :\\ &\quad \hat{\kappa_k} = \prod_{i \in I_S} (e(C', K_{k,i})/e(L_{k,i}, C_{\rho(i)}))^{\omega_i} = e(g,g)^{a\alpha_k s}\\ &\text{If } \hat{\kappa_1}^{\tau}\hat{\kappa_3} \neq d_1 \vee \hat{\kappa_2}^{\tau}\hat{\kappa_4} \neq d_2,\\ &\quad \text{then put } \hat{\kappa} = \perp, \text{else put } \hat{\kappa} = \hat{\kappa_1}.\end{aligned}$$

5.2 Security and Proof

Theorem 3 *If the Ostrovsky-Sahai-Waters* KP-ABKEM$_{cpa}$ *[11] is selectively secure against chosen-plaintext attacks and an employed hash function family* Hfam *has target collision resistance, then our* KP-ABKEM *is selectively secure against chosen-ciphertext attacks. More precisely, for any given PPT adversary* $\mathcal{A}$ *that attacks* KP-ABKEM *in the IND-sel-CCA game where decapsulation queries are at most* q_d *times, and for any small attribute universe* $\mathcal{U}$*, there exist a PPT adversary* $\mathcal{B}$ *that attacks* KP-ABKEM$_{cpa}$ *in the IND-sel-CPA game and a PPT target collision finder* $\mathcal{CF}$ *on* Hfam *that satisfy the following tight reduction.*

$$\begin{aligned}&\mathbf{Adv}^{ind\text{-}sel\text{-}cca}_{\mathcal{A},KP\text{-}ABKEM}(\lambda,\mathcal{U})\\ \leq &\mathbf{Adv}^{ind\text{-}sel\text{-}cpa}_{\mathcal{B},KP\text{-}ABKEM_{cpa}}(\lambda,\mathcal{U}) + \mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda) + \frac{q_d}{p}.\end{aligned}$$

Proof. We will omit the description of the proof because the proof goes analogously to the case of CP-ABKEM in Section 4.2. □

5.3 Encryption Scheme from KEM

It is straightforward to construct our encryption scheme KP-ABE from KP-ABKEM. The IND-sel-CCA security of KP-ABE is proved based on IND-sel-CPA security of the Waters *KEM* KP-ABKEM$_{\mathrm{cpa}}$.

Setup$(\lambda, \mathcal{U})$**.** The same as Setup of KP-ABKEM.

Encrypt$(\mathbf{PK}, \mathbb{A}, m)$**.** The same as Encap of KP-ABKEM except that Encrypt multiplies m by the blinding factor κ in the group $\mathbb{G}_T$. Encrypt returns $\mathrm{CT} = (C = m\kappa, \psi = (\psi_{\mathrm{cpa}}, d_1, d_2))$.

KeyGen$(\mathbf{MSK}, \mathbf{PK}, S)$**.** The same as KeyGen of KP-ABKEM.

Decrypt$(\mathbf{PK}, \mathbf{CT}, \mathbf{SK}_S)$**.** The same as Decap of KP-ABKEM except that Decrypt divides out C by the decapsulated blinding factor $\hat{\kappa}$. Decrypt returns the result $\hat{m}$.

5.4 Security and Proof

Theorem 4 *If the Ostrovsky-Sahai-Waters* KP-ABKEM$_{cpa}$ *[11] is selectively secure against chosen-plaintext attacks and an employed hash function family* Hfam *has target collision resistance, then our* KP-ABE *is selectively secure against chosen-ciphertext attacks. More precisely, for any given PPT adversary* $\mathcal{A}$ *that attacks* KP-ABE *in the IND-sel-CCA game where decryption queries are at most* q_d *times, and for any small attribute universe* $\mathcal{U}$*, there exist a PPT adversary* $\mathcal{B}$ *that attacks* KP-ABKEM$_{cpa}$ *in the IND-sel-CPA game and a PPT target collision finder* $\mathcal{CF}$ *on* Hfam *that satisfy the following inequality.*

$$\begin{aligned}&\mathbf{Adv}^{ind\text{-}sel\text{-}cca}_{\mathcal{A},KP\text{-}ABE}(\lambda,\mathcal{U})\\ \leq &2\Big(\mathbf{Adv}^{ind\text{-}sel\text{-}cpa}_{\mathcal{B},KP\text{-}ABKEM_{cpa}}(\lambda,\mathcal{U}) + \mathbf{Adv}^{tcr}_{\mathcal{CF},Hfam}(\lambda) + \frac{q_d}{p}\Big).\end{aligned}$$

Proof. We will omit the description of the proof because the proof goes in the same way as the case of CP-ABE in Section 4.4. □

6 Efficiency Discussion

First of all, we remark that our individual modification to attain CCA security is applicable when a Diffie-Hellman tuple to be verified is in the target group of a bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$. Especially, it is applicable even when an original CPA secure scheme is based on asymmetric pairing [19], $e : \mathbb{G}_1 \times \mathbb{G}_2 \to \mathbb{G}_T$. For example, the Type 3 version [19] of the Waters CP-ABE scheme [4] can be found in [20]. Detailed discussions and results on real implementations are found for the case of CPA-secure ABE schemes [21, 20]. We note here that the efficiency comparison below enables to guess the implementation results of CCA-secure ABE schemes via our modification.

We compare the efficiency of our CP-ABE with the original Waters CP-ABE$_{\mathrm{cpa}}$, and our KP-ABE with the original Ostrovsky-Sahai-Waters KP-ABE$_{\mathrm{cpa}}$. We also compare the efficiency of our schemes with the CCA-secure CP-ABE and KP-ABE schemes obtained by the generic transformation in [10]. Here the generic transformation [10] is considered in the case of a small attribute universe, the delegation type [10] and the Lamport one-time signature [22]. Table 1 shows these comparison. Note that a hash function is applied to generate a message digest of bit-length λ before signing by a secret key of the one-time signature. Note also, for simplicity, we evaluate the lengths and the amounts of computation below in the case that an access structure $\mathbb{A}$ is "all-AND" and an attribute map ρ is injective (i.e "single-use" that is opposed to "multi-use").

Table 1: Efficiency comparison of IND-sel-CCA secure ABEs ([10] and ours) with the original IND-sel-CPA secure ABEs [4, 11].

Scheme	L(PK)	L(SK$_S$)	L(CT)	C(Enc)	C(Dec)
Generic transform [10], CP-ABE	$+4\lambda^2(\mathbb{G})$	$+4\lambda^2(\mathbb{G})$	$+3\lambda^2$(bit)	$+2\lambda^2$exp.($\mathbb{G}$)	$+2\lambda^2$pair.(e)
Our individual modification (CP-ABE)	$+3(\mathbb{G}_T)$	$\times 4$	$+2(\mathbb{G}_T)$	$+4$exp.($\mathbb{G}_T$)	$\times 4$
Generic transform [10], KP-ABE	$+4\lambda^2(\mathbb{G})$	$+0$	$+3\lambda^2$(bit)	$+2\lambda^2$exp.($\mathbb{G}$)	$+2\lambda^2$pair.(e)
Our individual modification (KP-ABE)	$+3(\mathbb{G}_T)$	$\times 4$	$+2(\mathbb{G}_T)$	$+4$exp.($\mathbb{G}_T$)	$\times 4$

1) λ is the security parameter. (For instance, $\lambda = 224$ or 256.)
2) L(data) denotes the length of the data. C(algorithm) denotes the computational amount of the algorithm.
3) + and $\times$ mean the increment and the multiplier to the length or to the computational amount of the Waters CP-ABKEM$_{\mathrm{cpa}}$ and the Ostrovsky-Sahai-Waters KP-ABE$_{\mathrm{cpa}}$.
4) ($\mathbb{G}$), ($\mathbb{G}_T$) and (bit) mean that the lengths are evaluated in the number of elements in $\mathbb{G}$, elements in $\mathbb{G}_T$ and bits, respectively.
5) exp.($\mathbb{G}$) and pair.(e) mean the computational amount of one exponentiation in $\mathbb{G}$ and one pairing computation by the map e, respectively.

Our individual modification results in expansion of the length of a secret-key and the amount of decryption computation by a factor of four, while the length of a public-key, the length of a ciphertext and the amount of encryption computation are almost the same as those of the original CPA-secure schemes. In the case that the size of an attribute set is up to ($\frac{2}{3}$ of) the square of the security parameter λ, the amount of decryption computation of our CP-ABE and KP-ABE are smaller than those of the CP-ABE and KP-ABE obtained by the generic transformation [10], respectively.

7 Conclusion

We demonstrated direct chosen-ciphertext security modification for ABE in the standard model in the case of the Waters scheme (CP-ABKEM$_{\mathrm{cpa}}$, CP-ABE$_{\mathrm{cpa}}$) and the Ostrovsky-Sahai-Waters scheme (KP-ABKEM$_{\mathrm{cpa}}$, KP-ABE$_{\mathrm{cpa}}$). We utilized the Twin Diffie-Hellman Trapdoor Test of Cash, Kiltz and Shoup and the algebraic trick of Boneh and Boyen [13] and Kiltz [14]. Our modification worked for the setting that the Diffie-Hellman tuple to be verified in decryption was in the target group of the bilinear map. We compared the efficiency of our CCA-secure ABE schemes with the original CPA-secure ABE schemes and with the CCA-secure ABE schemes obtained by the versatile generic transformation.

Acknowledgment This work was supported by JSPS KAKENHI Grant Number JP15K00029.

References

[1] Amit Sahai and Brent Waters. Fuzzy identity-based encryption. In *Advances in Cryptology - EUROCRYPT 2005, 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Aarhus, Denmark, May 22-26, 2005, Proceedings*, pages 457–473, 2005.

[2] Vipul Goyal, Omkant Pandey, Amit Sahai, and Brent Waters. Attribute-based encryption for fine-grained access control of encrypted data. In *Proceedings of the 13th ACM Conference on Computer and Communications Security, CCS 2006, Alexandria, VA, USA, Ioctober 30 - November 3, 2006*, pages 89–98, 2006.

[3] John Bethencourt, Amit Sahai, and Brent Waters. Ciphertext-policy attribute-based encryption. In *2007 IEEE Symposium on Security and Privacy (S&P 2007), 20-23 May 2007, Oakland, California, USA*, pages 321–334, 2007.

[4] Brent Waters. Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization. In *Public Key Cryptography - PKC 2011 - 14th International Conference on Practice and Theory in Public Key Cryptography, Taormina, Italy, March 6-9, 2011. Proceedings*, pages 53–70, 2011.

[5] Allison B. Lewko, Tatsuaki Okamoto, Amit Sahai, Katsuyuki Takashima, and Brent Waters. Fully secure functional encryption: Attribute-based encryption and (hierarchical) inner product encryption. In *Advances in Cryptology - EUROCRYPT 2010, 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, French Riviera, May 30 - June 3, 2010. Proceedings*, pages 62–91, 2010.

[6] Hiroaki Anada and Seiko Arita. Short cca-secure ciphertext-policy attribute-based encryption. In *2017 IEEE International Conference on Smart Computing, SMARTCOMP 2017, Hong Kong, China, May 29-31, 2017*, pages 1–6, 2017.

[7] Ran Canetti, Shai Halevi, and Jonathan Katz. Chosen-ciphertext security from identity-based encryption. In *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*, pages 207–222, 2004.

[8] Dan Boneh and Xavier Boyen. Efficient selective-id secure identity-based encryption without random oracles. In *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*, pages 223–238, 2004.

[9] Xavier Boyen, Qixiang Mei, and Brent Waters. Direct chosen ciphertext security from identity-based techniques. In *Proceedings of the 12th ACM Conference on Computer and Communications Security, CCS 2005, Alexandria, VA, USA, November 7-11, 2005*, pages 320–329, 2005.

[10] Shota Yamada, Nuttapong Attrapadung, Goichiro Hanaoka, and Noboru Kunihiro. Generic constructions for chosen-ciphertext secure attribute based encryption. In *Public Key Cryptography - PKC 2011 - 14th International Conference on Practice and Theory in Public Key Cryptography, Taormina, Italy, March 6-9, 2011. Proceedings*, pages 71–89, 2011.

[11] Rafail Ostrovsky, Amit Sahai, and Brent Waters. Attribute-based encryption with non-monotonic access structures. In *Proceedings of the 2007 ACM Conference on Computer and Communications Security, CCS 2007, Alexandria, Virginia, USA, October 28-31, 2007*, pages 195–203, 2007.

[12] David Cash, Eike Kiltz, and Victor Shoup. The twin diffie-hellman problem and applications. In *Advances in Cryptology - EUROCRYPT 2008, 27th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Istanbul, Turkey, April 13-17, 2008. Proceedings*, pages 127–145, 2008.

[13] Dan Boneh and Xavier Boyen. Efficient selective-id secure identity-based encryption without random oracles. In *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*, pages 223–238, 2004.

[14] Eike Kiltz. Chosen-ciphertext security from tag-based encryption. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, pages 581–600, 2006.

[15] M. Choudary Gorantla, Colin Boyd, and Juan Manuel González Nieto. Attribute-based authenticated key exchange. In *Information Security and Privacy - 15th Australasian Conference, ACISP 2010, Sydney, Australia, July 5-7, 2010. Proceedings*, pages 300–317, 2010.

[16] Nuttapong Attrapadung, Benoît Libert, and Elie de Panafieu. Expressive key-policy attribute-based encryption with constant-size ciphertexts. In *Public Key Cryptography - PKC 2011 - 14th International Conference on Practice and Theory in Public Key Cryptography, Taormina, Italy, March 6-9, 2011. Proceedings*, pages 90–108, 2011.

[17] Nuttapong Attrapadung. Dual system encryption via doubly selective security: Framework, fully secure functional encryption for regular languages, and more. In *Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, pages 557–577, 2014.

[18] Moni Naor and Moti Yung. Universal one-way hash functions and their cryptographic applications. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing, May 14-17, 1989, Seattle, Washigton, USA*, pages 33–43, 1989.

[19] Steven D. Galbraith, Kenneth G. Paterson, and Nigel P. Smart. Pairings for cryptographers. *Discrete Applied Mathematics*, 156(16):3113–3121, 2008.

[20] Eric Zavattoni, Luis J. Dominguez Perez, Shigeo Mitsunari, Ana H. Sánchez-Ramírez, Tadanori Teruya, and Francisco Rodríguez-Henríquez. Software implementation of an attribute-based encryption scheme. *IEEE Trans. Computers*, 64(5):1429–1441, 2015.

[21] Ana Helena Sánchez and Francisco Rodríguez-Henríquez. NEON implementation of an attribute-based encryption scheme. In *Applied Cryptography and Network Security - 11th International Conference, ACNS 2013, Banff, AB, Canada, June 25-28, 2013. Proceedings*, pages 322–338, 2013.

[22] L. Lamport. Constructing digital signatures from a one-way function. Technical report, October 1979.

Appendix

A Proof of Lemma 2

The only one point to be complemented to the original proof (in [12]) is that even for any algorithm $\mathcal{A}$ with unbounded computational power, the statement holds. This is because, conditioning on the input fixed values (g, X_1, X_2), $\mathcal{A}$ only reduces the two-dimensional freedom $(r,s) \in \mathbb{Z}_p^2$ into the one-dimensional freedom $r \in \mathbb{Z}_p$ even if $\mathcal{A}$ correctly guesses the relation $s = rx_1 + x_2$. □

B Proof of Claim 1

For the index 3, we have implicitly set $t_3 = t_{\mathrm{cpa},3}(-\tau^*)$ and we get:

$$K_3 = g^{\mu_1} K_{\mathrm{cpa},3}^{-\tau^*} = g^{\mu_1}(g^{\alpha_1} g^{at_{\mathrm{cpa},3}})^{-\tau^*} = g^{\mu_1-\alpha_1\tau^*} g^{at_3} = g^{\alpha_3} g^{at_3},$$

$$L_3 = L_{\mathrm{cpa},3}^{-\tau^*} = (g^{t_{\mathrm{cpa},3}})^{-\tau^*} = g^{t_3},$$

$$K_{3,x} = K_{\mathrm{cpa},3,x}^{-\tau^*} = (h_x^{t_{\mathrm{cpa},3}})^{-\tau^*} = h_x^{t_3}, x \in S.$$

For the index 4, we have implicitly set $t_4 = t_{\mathrm{cpa},4}(\gamma_1\tau^*)$ and we get:

$$K_4 = g^{\mu_2-\gamma_2\tau^*} K_{\mathrm{cpa},4}^{\gamma_1\tau^*} = g^{\mu_2-\gamma_2\tau^*}(g^{\alpha_1} g^{at_{\mathrm{cpa},4}})^{\gamma_1\tau^*}$$

$$= g^{\mu_2-\gamma_2\tau^*} g^{\alpha_1\gamma_1\tau^*} g^{at_4} = g^{\mu_2-(\gamma_2-\alpha_1\gamma_1)\tau^*} g^{at_4} = g^{\mu_2-\alpha_2\tau^*} g^{at_4}$$

$$= g^{\alpha_4} g^{at_4},$$

$$L_4 = L_{\mathrm{cpa},4}^{\gamma_1\tau^*} = (g^{t_{\mathrm{cpa},4}})^{\gamma_1\tau^*} = g^{t_4},$$

$$K_{4,x} = K_{\mathrm{cpa},4,x}^{\gamma_1\tau^*} = (h_x^{t_{\mathrm{cpa},4}})^{\gamma_1\tau^*} = h_x^{t_4}, x \in S.$$

□

C Proof of Claim 2

Suppose that we are given a twin DH tuple $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$. Then, $d_i/e(C',g)^{\mu_i} = (e(g,g)^{\alpha_i})^{s(\tau-\tau^*)}, i = 1,2$. So, using the implicit relations (1), we have:

$$d_i = e(g,g)^{\alpha_i s(\tau-\tau^*)} e(g^s, g)^{\mu_i} = (e(g,g)^{\alpha_i(\tau-\tau^*)} e(g,g)^{\mu_i})^s$$

$$= (e(g,g)^{\alpha_i(\tau-\tau^*)} e(g,g)^{\alpha_i\tau^*+\alpha_{(i+2)}})^s = (e(g,g)^{\alpha_i\tau} e(g,g)^{\alpha_{(i+2)}})^s, i = 1,2.$$

This means that $(e(g,g),\ e(g,g)^{\alpha_1\tau} e(g,g)^{\alpha_3},\ e(g,g)^{\alpha_2\tau} e(g,g)^{\alpha_4}, e(C',g), d_1, d_2)$ is a twin Diffie-Hellman tuple.

The converse is also verified by the reverse calculation. □

D Proof of Claim 5

To reduce to the target collision resistance of an employed hash function family *Hfam*, we construct a PPT target collision finder $\mathcal{CF}$ that attacks *Hfam* using $\mathcal{A}$ as a subroutine. The construction is shown in Fig.2. (Note that the case Collision is defined in Fig.2.)

Note that the view of $\mathcal{A}$ in $\mathcal{CF}$ is the same as the real view of $\mathcal{A}$ until the case Collision occurs and hence the algorithm $\mathcal{A}$ works as designed.

To evaluate the probability in Claim 5, we consider the following two cases.

Case 1: the case that Abort ($\tau = \tau^*$) occurs in $\mathcal{B}$ in Phase 1. In this case, the target τ^* has not been given to $\mathcal{A}$. So $\mathcal{A}$ needs to guess τ^* to cause a collision $\tau = \tau^*$. Hence:

$$\Pr[\text{Phase 1} \wedge \Big(\bigwedge_{i=1}^{q_\mathrm{d}} \neg\textsc{Overlook}_i\Big) \wedge \textsc{Abort}] \leq \Pr[\text{Phase 1} \wedge \textsc{Collision}]. \tag{6}$$

Case 2: the case that Abort ($\tau = \tau^*$) occurs in $\mathcal{B}$ in Phase 2. In this case, if, in addition to $\tau = \tau^*$, it occurred that $\psi_{\mathrm{cpa}} = \psi^*_{\mathrm{cpa}}$ (and hence $C' = C'^*$), then it would occur that $\psi = \psi^*$. This is because the following two tuples are equal twin DH tuples by the fact that $\textsc{Overlook}_i$ never occurs:

$$(e(g,g), e(g,g)^{\alpha_1\tau} e(g,g)^{\alpha_3}, e(g,g)^{\alpha_2\tau} e(g,g)^{\alpha_4}, e(C',g), d_1, d_2),$$

$$(e(g,g), e(g,g)^{\alpha_1\tau^*} e(g,g)^{\alpha_3}, e(g,g)^{\alpha_2\tau^*} e(g,g)^{\alpha_4}, e(C'^*,g), d_1^*, d_2^*).$$

Hence both $S \in \mathbb{A}$ and $\psi = \psi^*$ would occur. This is ruled out in decapsulation query; a contradiction. So we have $\psi_{\mathrm{cpa}} \neq \psi^*_{\mathrm{cpa}}$; that is, a collision:

$$\psi_{\mathrm{cpa}} \neq \psi^*_{\mathrm{cpa}} \wedge H_\eta(\psi_{\mathrm{cpa}}) = \tau = \tau^* = H_\eta(\psi^*_{\mathrm{cpa}}).$$

Therefore, if $\textsc{Overlook}_i$ never occurs for each i, then only decapsulation queries for which $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ are certainly twin DH tuples have the chance to cause a collision $\tau = \tau^*$, as is the case in $\mathcal{CF}$. Hence we have:

$$\Pr[\text{Phase 2} \wedge \Big(\bigwedge_{i=1}^{q_\mathrm{d}} \neg\textsc{Overlook}_i\Big) \wedge \textsc{Abort}] \leq \Pr[\text{Phase 2} \wedge \textsc{Collision}]. \tag{7}$$

Given λ as input :
Set up
Initialize inner state;
Choose a polynomial size attribute universe $\mathcal{U}$ at random;
Get a target access structure $\mathbb{A}^* \leftarrow \mathcal{A}(\lambda, \mathcal{U})$;
Run $\mathbf{Setup}_{\mathrm{cpa}}(\lambda, \mathcal{U})$ to get $(p, \mathbb{G}, \mathbb{G}_T, g, e), \mathrm{PK}_{\mathrm{cpa}}, \mathrm{MSK}_{\mathrm{cpa}}$;
Get $(\kappa^*, \psi^*_{\mathrm{cpa}}) \leftarrow \mathbf{Encap}_{\mathrm{cpa}}(\mathrm{PK}_{\mathrm{cpa}}, \mathbb{A}^*)$;
Output ψ^*_{cpa};
Receive, in return, $\eta \leftarrow HKey(\lambda)$ and compute $\tau^* \leftarrow H_\eta(\psi^*_{\mathrm{cpa}})$;
Choose $\alpha_2, \alpha_3, \alpha_4 \leftarrow \mathbb{Z}_p$;
Put $\mathrm{PK} = (\mathrm{PK}_{\mathrm{cpa}}, e(g,g)^{\alpha_2}, e(g,g)^{\alpha_3}, e(g,g)^{\alpha_4}, \eta), \mathrm{MSK} = (g^{\alpha_1}, g^{\alpha_2}, g^{\alpha_3}, g^{\alpha_4})$;
Give PK to $\mathcal{A}$;
Phase 1
In the case that $\mathcal{A}$ makes a key-extraction query for $S \subset \mathcal{U}$;
Reply SK_S to $\mathcal{A}$ in the same way as **KeyGen** does using MSK;
In the case that $\mathcal{A}$ makes a decapsulation query for $(S, \psi = (\psi_{\mathrm{cpa}}, d_1, d_2))$;
Reply $\hat{\kappa}$ to $\mathcal{A}$ in the same way as **Decap** does using MSK;
If $\hat{\kappa} \neq \perp$ and $\tau = \tau^*$ (: call this case Collision)
then return ψ_{cpa} and stop;
Challenge
In the case that $\mathcal{A}$ makes a challenge instance query;
Using MSK, put $d_1^* = e(g^{\alpha_1}, C'^*)^{\tau^*} e(g^{\alpha_3}, C'^*), d_2^* = e(g^{\alpha_2}, C'^*)^{\tau^*} e(g^{\alpha_4}, C'^*)$,
$\psi^* = (\psi^*_{\mathrm{cpa}}, d_1^*, d_2^*)$;
Choose $\kappa \leftarrow KeySp(\lambda), b \leftarrow \{0,1\}$;
If $b = 1$ then put $\tilde{\kappa} = \kappa^*$ else put $\tilde{\kappa} = \kappa$;
Reply $(\tilde{\kappa}, \psi^*)$ to $\mathcal{A}$;
Phase 2
The same as in Phase 1;
Return
In the case that $\mathcal{A}$ returns its guess b^*;
Stop.

Figure 2: A PPT Collision Finder $\mathcal{CF}$ that attacks *Hfam* for the proof of Claim 5.

Taking a sum of both sides of (6) and (7), we get:

$$\Pr[\Big(\bigwedge_{i=1}^{q_{\mathrm{d}}} \neg \textsc{Overlook}_i\Big) \wedge \textsc{Abort}] \leq \Pr[\textsc{Collision}] = \mathbf{Adv}^{\mathrm{tcr}}_{\mathcal{CF}, Hfam}(\lambda).$$

□

E Proof of Claim 6

It is enough to prove that

When $(e(g,g), e(g,g)^{\alpha_1}, e(g,g)^{\alpha_2}, \hat{Y}, \hat{Z}_1, \hat{Z}_2)$ is a twin DH tuple,
$\hat{\kappa} = \hat{Z_1}^{1/(\tau - \tau^*)} = e(g,g)^{\alpha_1 s}$ holds.

This is deduced as follows:

$$\hat{\kappa} = (d_1 / e(C', g)^{\mu_1})^{1/(\tau - \tau^*)} = ((e(g,g)^{\alpha_1})^{s(\tau - \tau^*)})^{1/(\tau - \tau^*)} = e(g,g)^{\alpha_1 s}.$$

□

F Proof of Claim 7

A direct calculation with equalities (1) shows:

$$d_i^* = e(C'^*, g)^{\mu_i} = e(g,g)^{s^*(\alpha_i \tau^* + \alpha_{(i+2)})} = e(g,g)^{\alpha_i s^* \tau^*} e(g,g)^{\alpha_{(i+2)} s^*},$$
$i = 1, 2.$

Hence $\psi^* = (\psi^*_{\mathrm{cpa}}, d_1^*, d_2^*)$ is legitimate and correctly distributed. □

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 274-280 (2018)

www.astesj.com

ASTESJ

ISSN: 2415-6698

A Survey of Security Challenges in Internet of Things

Anass Sedrati*, Abdellatif Mezrioui

Telecommunication Systems, Networks and Services Lab, RAISS Team, INPT, Rabat, Morocco

ARTICLE INFO

Article history:
Received: 25 September, 2017
Accepted: 18 January, 2018
Online: 30 January, 2018

Keywords :
Internet of Things
Challenges
Security
Privacy
Lightweight

ABSTRACT

Internet of things (IoT) is an innovative technology subject to all kind of imaginary and science fictional solutions. Dreams and speculations are still possible about it. A technology combining real life objects and virtual life (Internet) is indeed a fertile pitch of fantasy and original ideas. However, IoT has in practice to face several challenges to ensure its function and operability in a near future. This paper defines first some technical challenges of IoT today, before focusing on security-related ones via a layered architecture of IoT that we suggest. Finally, a number of actions and required future work is presented to enhance IoT security (Privacy, Lightweight crypto, etc.).

1. Introduction

Nowadays, billions of people are active using Internet for all kinds of purposes on a daily basis. People send in fact emails, use social networks, share voice and image, transfer money, watch events, and perform many more actions with it. It is estimated that by 2020, there will be 50 to 100 billion devices connected to Internet [1]. If what is happening now was difficult to conceive 20 years ago, one can easily imagine that future will be as unpredictable, if not even more.

In this context, even Internet itself is set to change from its classical network infrastructure to a more flexible one: The Internet of Things (IoT). IoT will allow most objects to be connected to Networks and interact in different scales. This opens doors to new applications in all domains one can think about. A new way of living and working is emerging by embedding electronics into everyday physical objects. IoT is the next step of the development of communication tools. As a new technology allowing many "things" to be connected for the first time ever, IoT marks a clear difference with the classical Internet where only given devices could do so. This difference is the driver of this article. Given the specificity of IoT and the uniqueness of the "things" it involves, technologies used in Internet might be incompatible in many aspects. This incompatibility is the new challenge facing IoT, affecting many areas, mainly security. To have a functional and secure IoT technology in the future, issues as sensors/actuators and privacy should be looked into and solved. The aim of this article is to first summarize the challenges of IoT nowadays, before focusing then about our interest area: security issues that are facing the Internet of Things.

The article is organized as follows. In Section 2, we will introduce IoT and define it along with its main related concepts. In Section 3, we will list the differences between IoT and the "traditional" Internet. This comparison will lead us to Section 4, where IoT challenges are presented. Section 5 describes a model of IoT Architecture on which we are going to argue. This model will help us to identify security challenges. Section 6 will be a continuation of the previous Section, highlighting mainly security issues, and describing them a little bit more in detail. In Section 6, we will also refer to our model architecture when detailing security challenges, and link each security challenge to its place in the model. Section 7 will finally be the conclusion and an opening to the future and new work that research could dig into.

2. Internet of Things

Formally defined, the Internet of Things is a link between "objects" of the real world with the virtual world, thus enabling anytime, anyplace connectivity for anything and not only anyone. It refers to a world where physical objects and beings, as well as virtual data and environments, all interact with each other in the same space and time" [2]. According to this definition, one can already note the complexity of the transition from an Internet used for interconnecting end-user devices to an Internet used for interconnecting physical objects that communicate with each other

*Corresponding Author: Anass Sedrati, Email: sedrati@inpt.ac.ma

and/or with humans [3]. IoT combines anything in many thinkable ways. Through its definition, IoT suggests different areas of application: From smart cities to transport or health, all domains are expected to enable IoT in different extends. Agriculture, industry, aeronautics, and even daily life would use it. One can e.g. call his self-driving car to pick him up at the door, and warm up coffee at the work desk while still sitting in traffic jam. Clothes can also with help of weather forecast "decide" how thin they shall be when a person is to wear them. From the most basic and common applications to the most complex and fanciest ones, IoT will be present in our lives very soon. In order to enable IoT, each "thing" should have the ability to support three pillars [3]: (i) The object should be identifiable (anything identifies itself), (ii) The object should be able to communicate (anything communicates), and should be able to (iii) interact (anything interacts). If they fulfill those pillars, "things" can be considered as "smart objects".

In addition to the three pillars, each "thing" should contain a set of technical components that will enable IoT technology into it. Two essential components of IoT are sensors and actuators. A sensor is a little electronic component able to fulfill the three pillars defined above. In order to make a thing "smart" and able to realize the three pillars, a sensor or an actuator should be attached to it. Through this linking, the thing will be able to send and receive data, and become part of the IoT. A set of sensors form a network called "Sensor network" where they can interact with each other or with Internet. When the networks are wireless, they can be referred to as WSNs (Wireless Sensor Networks). An actuator is an element that converts energy into motion. There are three types of actuators: Electrical, hydraulic and pneumatic [4]. The interest of actuators in as IoT context is that they allow smart objects to trigger actions having an effect on the physical reality [3]. Actuators can be combined with sensors and connected in a network structure called SANET (sensor/actuator network).

To summarize, we can then say that IoT is the network gathering all "things" having a sensor/actuator, and connected with each other in all possible ways and forms. These objects can exchange data to different extends depending on variables that we will see later on.

3. Technical Differences Between IoT and Internet

As we have seen in the definitions, IoT can be considered as the next generation of Internet, allowing all sorts of objects to be connected. The first difference is then that not only specific devices can access networks, but basically all of them. The only condition is having a sensor/actuator that can communicate and support the three pillars (being identifiable, communicate and interact). A list of major differences between IoT and Internet will contain:

- Sensors/Actuators: Because "things" are not initially set to be connected, they have to be implemented with sensors/actuators. In classical Internet, devices (PC, Smart Phone, TV…) have a complicated electronic system. In IoT, things have a main role that is not always technological. Clothes role is to keep warm, and roofs' is to protect houses from rain and snow. When these objects will be connected, sensors/actuators should not constitute a major part of them (because they still have to fulfill their initial aim). These technological elements must however be present in order to make the objects labelled as "smart". A solution is to have smaller sensors and actuators. This means that they will be limited in resources and capacity, unlike devices used in Internet. Moreover, Internet Enabled devices do not have power consideration, and use chargers.
  If sensors are also used in classical Internet such as in IP cameras, the difference is that in IoT, sensors have usually to be on low-power. Their charging (or being self-charging) is still a hanging issue. This difference creates many challenges that will be discussed in the next section.
- Autonomous "Things": In IoT, Things are expected to be more autonomous than our usual devices. Some objects should be able to perform a number of duties themselves, and to communicate with each other without human interaction. But not only that, some specific cases in IoT imply that things are directly connected with each other in their network (Fridge and a car in the smart home network), which creates an extra complexity knowing that no human intervention should be in this network.
- Difference in nodes [5]: IoT can be composed of Radio-frequency Identification (RFID) and WSN nodes, whose resources are limited, while the Internet is composed of PC, servers, smart phones whose resources are rich. This means that combinations of complex algorithms can be used in Internet, while IoT is limited in this aspect. Then, other alternatives for security need to be found for IoT.
- Heterogeneity: IoT involves very heterogeneous objects which can have different standards. In Internet, data formats and standards are similar even in different operative systems. Managing this heterogeneity in contents and formats is an important milestone for IoT enabling.

The difference between Internet and IoT can be summarized in the Figure 1 in the next page. Figure 1 shows the evolution of networks with respect to the type of connected devices and objects. Internet of Things is gathering not only traditional "technical" devices, but extends even to daily life objects such as cars, fridges or houses. It can be considered as the evolution of Internet towards a connected daily-life objects.

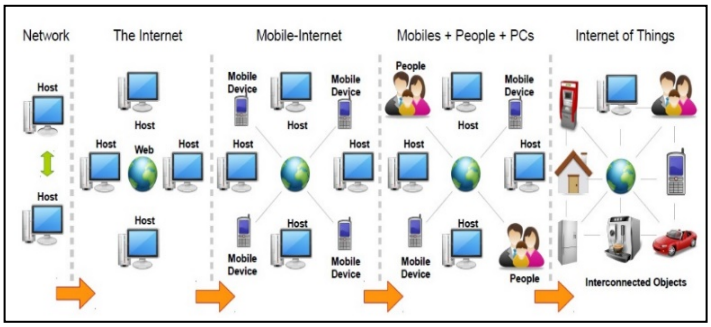


Figure 1. Evolution from Internet to Internet of Things [2].

Given the differences between IoT and Internet that we saw in this section, a preliminary list of what IoT needs to support can be established [3,5]:

- Manage devices heterogeneity
- Scalability: Naming, addressing, managing information…

- Spectrum Availability: Taking into consideration that the number of IoT connected devices will be much higher than in current Internet.
- Self-organization capabilities: As "things" will be autonomous and make decision by themselves in some situations.
- Context awareness: A device can have different modes and roles depending on the context. The device should be aware about it by its own.
- Security and Privacy: As IoT has to find new ways of securing, different from the complex algorithms used in the Internet. IoT devices have in fact less resources (power, CPU, etc.).

4. Internet of Things Challenges

The differences that we have seen between IoT and classical Internet in the previous Section are the main reason of today's IoT challenges. Internet is in fact an established technology, and a number of its critical issues were already solved. Those challenges can be the same ones facing IoT. The problem is that IoT with its differences in components and way of working creates a new "versions" of these challenges. The difference with Internet implies in fact to find new innovative ways of solving, instead of the "traditional" ones, which only work with Internet. In order to be operational in the future, IoT needs to solve the most critical of them at least.

Considering what we have seen about IoT, we can already list the following challenges:

- Sensors/Actuators: As all items (things) will be connected to Internet, they will need a tool to link them to the network: It is the sensors/actuators. Those elements implemented in all sorts of "things" (clothes, walls, fridge…) must be ready to work at any time. Internet is indeed working real-time, and information ought to arrive at almost the same time it is sent. This scenario supposes that all connected "things" have charged sensors at every moment to allow their discovery and be able to send and receive data (resp. actuators in order to be able to perform actions). The challenge is then an energetic one. Low-power wireless sensors which do not need battery replacement over their lifetimes are needed. How will those sensors be charged? Are they sustainable? Energy and power management is a principal issue within the IoT research area.
- Identification: In IoT, not all "things" have the same role. In the example of a smart house, a Fridge and the security camera both must be connected, but they do not have the same role and access specificities. This means that Objects should be identifiable in order to allow each one of them to perform its own duties. Identification can be either by being part of a certain class (desk objects, kitchen tools…) or by unique identification. Identification is a notable challenge in IoT. Miorandi et al. [3] have suggested identifying objects in IoT in two ways: "The first one is to physically tag one object by means of RFIDs, QR code or similar (…) returning an identifier that can be looked up in a database for retrieving the set of features associated to it. The second possibility is to provide one object with its own description: if equipped with wireless communication means it could communicate directly its own identity and relevant features. It is however important to mention that description is not enough to make an object unique. Other elements as ownership have to be added and updated in order to preserve the privacy aspect. This is particularly relevant in scenarios regarding two cars of the same brand parked aside.
  The two approaches cited above are not mutually exclusive and can complement each other".
  The identification problem is not entirely solved and is always holding.
- Scalability: Given the expected huge number of objects that will be connected with IoT, network and frequency have to anticipate the enormous flow coming in soon, by scaling the network at different levels. Addressing is one example of scalability issues in IoT. The standard commonly used today with internet is the Internet Protocol (IP), and the biggest chances will be that addressing in IoT will also be in this protocol. Even the IPv6 protocol, a candidate for addressing, can face this challenge. But in This situation other questions are raised such as: Should each "thing" have an address at every moment? Should it be allowed a temporary/permanent address? What is the best scheme to identify each "thing" in the IoT?
  Scalability in not only related to addressing. Other challenges regarding the size can be about data and networking, information and knowledge management, and even service provisioning and management [3].
- Heterogeneity: IoT involves different "objects" ranging from the smallest chip to the big airplanes and buildings. It is not sure that all those items use the same set of protocols and data formats, due to their different capacities and size. However, if we want all those objects to communicate with each other, standards need to be implemented. Standardization does not only apply to addressing, but even to other areas in the IoT. One of the important institutions working to solve it is the IEEE. Challenges face indeed packets that will be routed through different sorts of networks. All those networks must follow the same norms and specifications to be synchronized and understand each other. Standards of IoT must cover nowadays many areas such as security, privacy, architecture and communications.
- Governance: The smarter "things" become, the most autonomous they are, and thus the more governance is needed. The question of regulation and how users will be protected are important in that matter. Which organizations will care about law enforcement for IoT? Should new ones be created? How should personal data be protected? Who can see what? How to prevent third party apps from accessing this data?
- Security: Security is a major and critical issue in IoT. Therefore, it will be separately presented in detail in the next two Sections.

5. An IoT reference Model

In order to analyze security aspects of IoT, a reference model would be of a good use. Many IoT reference models have been widely discussed in academic publications and these reference models distinguish different levels [6-8]. Providing detailed descriptions of all these IoT reference models is beyond the scope of this paper. But in order to analyze and summarize IoT security issues level-by-level we will use a simple three-level reference

model as depicted in Figure 2. This architecture is our starting point and it will be the architecture that we will consider in this article.

Thus, as shown in Figure 2, IoT can be broken down into three major layers: Devices collect data, gateways and communication units relay the data collected, applications and services analyze the data, and take actions. This architecture highlights also some security aspects that are related to the three main layers:

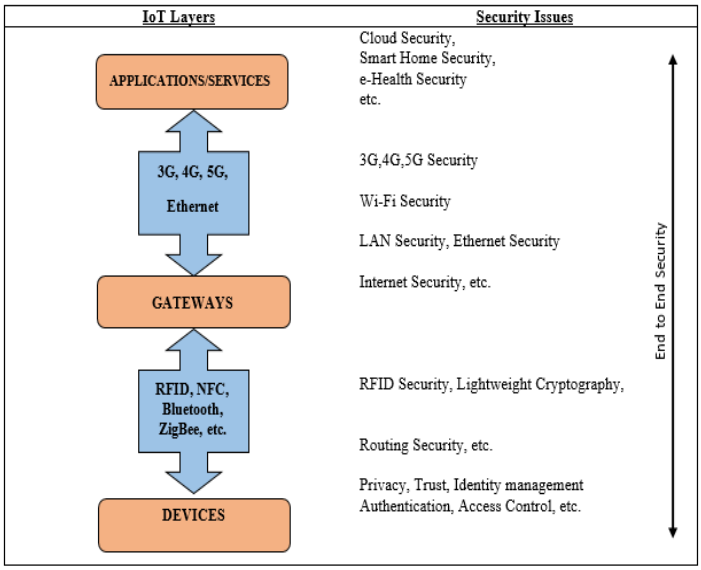


Figure 2. A reference Security Architecture for IoT.

- **Devices**: The Devices Layer is divided between nodes and network. It includes Radio-frequency Identification (RFID) security and Wireless sensors network (WSN). Wireless Sensor Network, which we have defined in Section 2, face in the IoT context many issues that we will define, such as heterogeneity, Cryptographic algorithms or Node trust management.
- **Gateways**: It is the layer responsible for the transport of data and the transmission of commands between the first and third layer (applications and services). This layer gathers security tools and protocols that are responsible of transporting, in a secure manner, data in 3/4/5G, WIFI and LAN/WAN networks.
- **Applications/Services:** This layer provides user applications and services. They can be accessed via cloud computing. These IoT applications and services are subjects to many attacks. Usual attacks to be stopped are Denial of Service (DDoS) attacks and Third-Party attacks. Security should be guaranteed into IoT applications (such as smart home or intelligent traffic), and platforms for support, such as cloud computing should be monitored and secured.

6. Security issues in IoT

6.1. IoT Security Requirements

Traditionally, security requirements can be broken down into three main categories: (i) confidentiality, (ii) integrity, and (iii) availability, referred to as the CIA-triad [9]. Confidentiality means limiting the access of certain information only to authorized parts. It is necessary in the IoT context especially regarding applications where information is critical, e.g. Health, finance. Integrity ensures that the received commands and information have not been changed. In case of an error, dramatic consequences could happen, mainly in Things working closely with human lives. Finally, availability ensures that all system services are available, when requested by an authorized user. These basic security principles (Confidentiality, Integrity and Availability) must be ensured by services and mechanisms adapted to the field of IoT.

Table 1. Security threats facing objects by lifecycle [15].

	Manufacturing	Installation	Operation
Applications / Services		Eavesdropping& Man-in-the-middle	Eavesdropping& Man-in-the-middle
Gateways		Eavesdropping& Man-in-the-middle	DDos Attack Routing attacks
Devices	Device cloning	Substitution	DDoS attack Privacy threat Extraction of security parameters

When investigating IoT security, there are also important requirements that need to be taken into consideration such as privacy, identity management, trust, End-to-End Security, authentication and access control. They are the main security issues which are to be addressed [10-13]. IoT implies “things”. They are the principal component of this technology. Heer et al. [14] have defined the lifecycle of a thing. Each object has three cycles which are: Manufacturing, installation and operation and there are vulnerabilities and threats facings objects during their lifecycle. Table 1 below summarizes different threats affecting the objects during each cycle with respect to the architecture we have defined in Figure 2.

On another side, when talking about security, an important related notion must be also introduced; Context awareness. In order to define different degrees of security, the context is an important factor to be taken into consideration. The object might indeed be secure in a given context, but exposed to threads in another. As objects are set to be autonomous in IoT, they should be aware of their context themselves. This is context awareness. Formally defined “a system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user task “[2]. When relating to security, we must think that as “smart” objects, security standards for things should vary depending on the context, as those objects are supposed to be context-aware. It is also important to know that context-awareness raises privacy issue. A device that is in fact aware of its context possesses a number of information that can

be valuable. Moreover, context-awareness poses an energetic challenge since the device needs to have constant and updated information, which affects its energy consumption.

Let us now study in deep those issues with the help of the architecture that was introduced in the previous section (Figure 2).

6.2. Applications/Services security issues

Attacks on the cloud

Clouds are nowadays mobile databases. Moving to cloud technology is more and more frequent, and even IoT follows this trend. Many IoT solutions have already connected platforms to the cloud such as Bolt [16] providing data management for the Lab of Things (LoT) [17] by using Amazon S3 or Azure for Data storage. Ninja Sphere [18] and Smart-Things Hub [19] do also deploy their solution of the cloud. It seems in fact easier and more practical to have a database on the cloud for IoT applications, in order to access data wherever the objects are. However, clouds are more subject to various kinds of attacks, given their centralizing role.

Many critical reports (Internet of Crappy Things, Internet of Fails, FTC technical reports) have emphasized security in IoT spaces [20]. By centralizing IoT in clouds, and given the present vulnerabilities, the risk of accessing information that can be confidential is even bigger. This might be of negative consequences especially if it is on a big scale or sensitive information (military, financial, etc.).

Service Interruption

DDoS attacks are a very common issue on the Internet, but they need also to be solved in an IoT context. Sonar et al [21] have surveyed these attacks on IoT. DDoS attacks on IoT can affect different layers, (perception, network or application). These attacks cause service interruption, as the cloud server becomes unresponsive, but they can also end on extremely slow response that leads to a deterioration of service quality.

Service interruption in IoT can also happen due to different other factors. Some of these interruptions have been investigated, such as trojans in IoT [22] and various viruses (Stuxnet [23]).

Third party attacks

Weber [24] defines Privacy Enhancing Technologies (PET) as technologies developed in order to achieve information privacy goals. They include Virtual Private Networks (VPN), Transport Layer Security (TLS), DNS Security Extensions (DNSSEC) or Peer-to-Peer (P2P). These technologies are not completely protective in the case of IoT. Certain situations of IoT such as positioning do still present privacy problems [25]. Some malicious third party can in fact access a user's location information and use it for malicious activities. Location information can be extracted for example from sensors and actuators while communicating.

Not only location can be hacked, but more sensitive information: Income or health status…As IoT chips can rely on RFID, they can also be subject to attacks getting private information.

6.3. Gateway related security issues

The gateway is related to the network and how data is transported. Network security is an essential part in protecting data in IoT. Regarding transport aspect, IoT is using the same standards as Internet, issues will then be similar. It presents the following security issues:

Wi-Fi security

The most used wireless standard is Wireless Fidelity (Wi-Fi). Wi-Fi users can encounter different kind of security threads, such as phishing websites. DDoS attacks can also happen on Wi-Fi and flood the network [21].

Access is also important for Wi-Fi. Not everybody can access a Wi-Fi network, but only those having the password. The problem is as we saw that in IoT not all "objects" have a user interface, or keyboard. How could we write a password in a smart T-shirt that only contains a chip with sensor?

Moreover, if the Wi-Fi network is not properly securitized, there is a risk that a connected object in a given network can access to other objects in the same network. It is a very frequent issue on Internet nowadays, and is certainly threat even for IoT security.

Other standards

Third Generation (3G) networks present also certain vulnerabilities. They are for instance subject to DDoS attacks, phishing attacks and identity attacks [5].

6.4. Devices security issues

Heterogeneity in technologies

As we have seen earlier, IoT is a technology that allows all sorts of objects to connect. By doing so, heterogeneity issue is raising even in Security matters. There is for example no uniform international encoding standard for RFID tag; this can create access problems or errors in reading process for the user [2]. Not only tags are different, but even data itself. Data can come with different or even incompatible formats. This can result in data loss or destruction, causing privacy exposure. There should be a process of unification of formats and protocols in order to guarantee a better security in IoT.

Encryption

Encryption is one of the fundaments of the modern internet security. Information cannot be send directly (or it can be intercepted on the way and be potentially misused from malicious parties). Encryption is the part where information is coded with the help of a key into another series of characters. Only parts with

the key can retrieve the original message and read it. Unfortunately, the classical encryption algorithms and standards cannot be applied for the IoT.

Many "things" do not support those algorithms that require a big memory. The "things" were in fact not designed in the first term to be connected, but rather to fulfill their natural function (clothes, walls, fridges...). This does not mean that those objects do not need to be protected, on the contrary, it is necessary. One of the suggestions given to encrypt data in IoT is then Lightweight cryptography.

The term "lightweight" should not be mistaken with weak (in terms of cryptographic protection), but should instead be interpreted as referring to a family of cryptographic algorithms with smaller footprint, low energy consumption, and low computational power needs, which will resolve both energy and security challenges.

Examples of uses of lightweight cryptography in IoT are authentications schemes [26-27] or device management [28-29]. Lightweight cryptography contains different sorts of algorithms that can be used, all of them are under studies, one can give as examples: Symmetric ciphers for lightweight crypto, asymmetric ciphers and homomorphism.

Trust Management

Trust is an important and necessary criterion in all transactions. It is defined as "the measurement of the belief from a trusting party point of view (trustor) with respect to a trusted party focused on a specific trust aspect that possibly implies a benefit or a risk" [30].

Technology and IoT are not an exception, and exchanges in IoT should all come from and to trusted parts. However, heavy encrypting and complex computing are not possible in IoT. This means that the trust system should be simple, but efficient. During the authentication period, user should easily be able to login, while having a secure system in front of him. This can be difficult as it looks as a paradox.

A secure system is usually complicated and difficult to use for a novice user, while a four number PIN code is rather easy to break and presents a weak security model. Then, one challenge research is highlighting is to invent new rich authentication mechanisms, that can be used for IoT, having a better security, but also being simple for use for any costumer and supported by the sensors.

Secure Routing Protocols

As IoT is limited in power and computing abilities, classical routing protocols can unfortunately not be used. One of the most important challenges is to design new secure routing protocols for Wireless Sensor Networks, as routing is a vital part of networking, and attacks toward a weak routing protocol can lead to the whole network collapse.

6.5. IoT Security Threats and Attacks

In Section 5, we have defined a reference model to analyze IoT security aspects. The model proposed to break down IoT into three different layers: Devices, Gateways and Applications / Services. In the current section, we have separately presented threats and attacks related to each specific layer.

Table 2. IoT Security Attacks by layer.

	Device Layer	**Gateway Layer**	**Applications / Services Layer**
Threats / Attacks	Heterogeneity in technologies Encryption Turst Management Secure Routing Protocols Node Tampering [31] Physical Damage Social Engineering [32]	Wi-Fi related Attacks 3G related Attacks RFID Attacks (Spoofing , Cloning, Unauthorized Access) Man in the Middle [33] Sybil Attack [34] Sinkhole Attack [35]	Attacks on the Cloud Service Interruption (DDos Attacks, Virus, Trojan, etc.) Third Party Attacks Spyware, Adware [36] Side Chanel Attacks [37] Cryptanalysis Attacks [32]

In order to summarize our work, we present Table 2 below that gathers different attacks and security threats facing IoT through its different layers. The table includes also some attacks that were not presented in this work. Due to constraints, we have in fact chosen to limit us to a number attacks for this article. Readers wishing to deepen their knowledge about the other attacks can investigate the references attached to them in the table.

7. Conclusions and future work

Internet of Things is for sure an amazing and exciting area, with many challenges ahead. We have first defined formally IoT and discovered its specificities. Then we detailed its main differences with the classical Internet. Understanding those differences is the key to specify the areas of challenges that IoT will face. After a general presentation of those challenges, we went into a suggested reference model to analyze security of IoT. The architecture of this model helped us to exhibit security issues that are hanging until now.

In this article, the aim was to give a general image of IoT, with a special focus on security. Our planned future work will be on the Lightweight security solutions. As objects in IoT do not support complex computing, and as cryptography is still important to securitize data, we would like to study on the future Lightweight cryptography algorithms. This is in fact an ongoing research issue that can be of high interest for the scientific community, and we would like to participate with our little contribution in this huge project of IoT.

References

[1] C. Perera, A. Zaslavsky, P. Christen, D. Georakopoulos, "Context Aware Computing for The Internet of Things: A Survey". IEEE Communications Surveys & Tutorials (May 2013).

[2] H. Sundmaeker, P. Guillemin, P. Friess, S. Woelffl, "Visions and challenges for realizing the internet of things, Cluster of European Research". Projects on the Internet-of-Things (CERPIoT, 2010).

[3] D. Miorandi, S. Sicari, F. De Pellegrini, I. Chlamtac, "Internet of Things: Vision, Applications and Research Challenges". Volume 10, Issue 7, pp. 1497-1516, Ad Hoc Networks, (September 2012).

[4] S. Madakam, R. Ramaswamy, S. Tripathi "Internet of Things (IoT): A Literature Review". Journal of Computer and Communications, 2015, 3, pp. 164-173

[5] Q. Jing, A.V. Vasilakos, J. Wan, J. Lu, D. Qiuv, "Security of the Internet of Things: perspectives and challenges". Volume 20, issue 8, pp. 2481-2507. Wireless Networks (November 2014).

[6] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," Computer Networks, vol. 54, no. 15, pp. 2787–2805, 2010.

[7] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Generation Computer Systems, vol. 29, no. 7, pp. 1645–1660, 2013.

[8] "The Internet of Things reference model." CISCO, 2014. [Online]. Available: http://cdn.iotwf.com/resources/71/IoT_Reference_Model_White_Paper_June_4_2014.pdf

[9] A.M. Nia, N.K. Jha, "A Comprehensive Study of Security of Internet-of-Things". Volume: PP, Issue: 99, IEEE Transactions on Emerging Topics in Computing (September 2016).

[10] N. Parikshit Mahalle, "Identity Authentication and Capability Based Access Control (IACAC) for the Internet of Things », Journal of Cyber Security and Mobility, Vol. 1, 309–348. 2013.

[11] Guanglei Zhao, "A novel mutual authentication scheme for Internet of Things". In Proceedings of 2011 IEEE International Conference on Modelling, Identification and Control (ICMIC), pp. 563–566, 26–29 (June 2011).

[12] C. Mayer. "Security and privacy challenges in the IoT". WowKivs, Electronic Communications of the EASST, Volume 17, Germany (2009).

[13] "Internet of Things: privacy and security in a connected world", US Federal Trade Commission, Staff report, 2015. https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf

[14] T. Heer., Garcia-Morchon, O., Hummen, R., Keoh, S.L., Kumar, S.S. and Wehrle, K., 2011. Security Challenges in the IP-based Internet of Things. *Wireless Personal Communications*, *61*(3), pp.527-542.

[15] C. Lu, "Overview of Security and Privacy Issues in the Internet of Things", May 2014

[16] T. Gupta, R.P. Singh, Mahajan, A. P. J. J. R. Bolt: Data management for connected homes. In Proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14) (2014), pp. 243–256.

[17] A. Brush, E. Filippov, D. Huang, J. Jung, R. Mahajan, F. Martinez, K. Mazhar, A. Phanishayee, A. Samuel, J. Scott, Et al." Lab of things: a platform for conducting studies with connected devices in multiple homes". In Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication (2013), ACM, pp. 35–38

[18] Ninja Blocks. https://ninjablocks.com/.

[19] SmartThings. http://www.smartthings.com/

[20] B. Zhang, N. Mor, J. Kolb, D.S. Chan, K. Lutz, E. Allman, J. Kubiatowicz, (2015, July). " The Cloud is Not Enough: Saving IoT from the Cloud". In *HotCloud*.

[21] K. Sonar, H. Upadhyay (2014). A survey: DDOS attack on Internet of Things. *International Journal of Engineering Research and Development*, *10*(11), 58-63.

[22] C. Liu, P. Cronin, C. Yang (2016, January). "A mutual auditing framework to protect IoT against hardware Trojans". In *Design Automation Conference (ASP-DAC), 2016 21st Asia and South Pacific* (pp. 69-74). IEEE.

[23] M.M.M. Dakhani, M.Z.A.I. Dakhani,(2017). Another Way to Deal with Research Privacy in the IOT: Threats and Assaults on IOT and its Solutions.

[24] R. H. Weber,"Internet of Things – New Security and Privacy Challenges", Computer Law & Security Report (January 2010).

[25] M. Elkhodr, S. Shahrestani, H. Cheung (2013, April). The Internet of Things: vision & challenges. In *TENCON Spring Conference, 2013 IEEE* (pp. 218-222). IEEE.

[26] M.A. Jan, P. Nanda, X. He, R-P. Liu (2016). A lightweight mutual authentication scheme for IoT objects. *IEEE Transactions on Dependable and Secure Computing (TDSC), "Submitted*.

[27] J.Y. Lee, W.C. Lin, Y.H. Huang (2014, May). A lightweight authentication protocol for internet of things. In *Next-Generation Electronics (ISNE), 2014 International Symposium on* (pp. 1-2). IEEE.

[28] T. Perumal, S.K. Datta, C. Bonnet (2015, October). IoT device management framework for smart home scenarios. In *Consumer Electronics (GCCE), 2015 IEEE 4th Global Conference on* (pp. 54-55). IEEE.

[29] Y. Jin, M. Tomoishi, N. Yamai (2017, July). A Secure and Lightweight IoT Device Remote Monitoring and Control Mechanism Using DNS. In *Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual* (Vol. 2, pp. 282-283). IEEE.

[30] R. Neisse, M. Wegdam, M. Van Sinderen "Trust management support for context-aware service platforms". In: User-centric networking, lecture notes in social networks. Springer international; 2014. p. 75-106.'

[31] A. Perrig, J. Stankovic, and D. Wagner, "Security in wireless sensor networks." Communications of the ACM 47, no. 6 (2004): 53-57.

[32] I. Andrea, C. Chrysostomou, G. Hadjichristofi, 2015, July. Internet of things: Security vulnerabilities and challenges. In *Computers and Communication (ISCC), 2015 IEEE Symposium on* (pp. 180-187). IEEE.

[33] R. P. Padhy, M. R. Patra, and S. C. Satapathy, "Cloud Computing: Security Issues and Research Challenges." International Journal of Computer Science and Information Technology & Security (IJCSITS) 1, no. 2 (2011): 136-146.

[34] J. Newsome, E. Shi, D. Song, and A. Perrig, "The sybil attack in sensor networks: analysis & defenses." In Proceedings of the 3rd international symposium on Information processing in sensor networks, pp. 259-268. ACM, 2004.

[35] V. Soni, P. Modi, and V. Chaudhri, "Detecting Sinkhole attack in wireless sensor network." International Journal of Application or Innovation in Engineering & Management 2, no. 2 (2013).

[36] Y.J. Jia, Q.A. Chen, S. Wang, A. Rahmati, E. Fernandes, Z. M. Mao, A. Prakash, S.J. Unviersity, 2017. ContexIoT: Towards Providing Contextual Integrity to Appified IoT Platforms. In *Proceedings of the 21st Network and Distributed System Security Symposium (NDSS'17)*.

[37] S. Mangard, E. Oswald, and T. Popp, Power Analysis Attacks: Revealing the Secrets of Smart Cards. Springer Verlag, 2007.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 281-287 (2018)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

Stabilization of constrained uncertain systems by an off-line approach using zonotopes

Walid Hamdi*, Wissal Bey, Naceur Benhadj Braiek

Laboratory of Advanced Systems (L.A.S), Polytechnic School of Tunisia, Carthage University, BP 743,2078 La Marsa, Tunisia

ARTICLEINFO

Article history:
Received: 30 November 2017
Accepted: 07 January 2018
Online: 31 January 2018

Keywords:
Stabilization
Zonotopic invariant sets
Model Predictive Control
Uncertain systems

ABSTRACT

In this paper, stabilization of uncertain systems was established using zonotopic sets. The obtained state feedback control laws are computed by an off-line approach reducing computational burdens. The resolution of a robust model predictive control (MPC) allows computing a sequence of state feedback control laws corresponding to a sequence of zonotopic invariant sets. The implemented control laws are then calculated by linear interpolation between the state feedback gains corresponding to the nested pre-computed zonotopic sets. The proposed interpolation with the use of zonotopic sets achieves better control performances.

1 Introduction

Model predictive control (MPC) is one of the most successful techniques of advanced control in the process industry. Thanks to the recent developments of the underlying theoretical framework, MPC has become a mature control technique able to provide controllers ensuring stability, robustness, constraint satisfaction and tractable computation for linear and nonlinear systems [1]. The MPC is can be made in the context of representation in state variables [2]. This not only make use of existing theorems and results in the state space theory, but also facilitates the extension of the theory of model predictive control to more complex cases such as systems with stochastic disturbances, noise on measured variables or multivariable control. For nonlinear uncertain systems, explicitly modeling of the uncertainty is essential [3].

For modeling uncertain systems, it is very important for MPC to be more robust [2]. Important areas in MPC that have recently seen significant theoretical and implementational progress include robust and stochastic MPC as well as efficient computations for MPC via convex and reliable real-time optimization [4].

Although these MPC schemes have remarkable performance and good theoretical properties, there is a hard computational burden due to the min-maximization of the optimization problem, especially in the presence of the system nonlinearity. The other is to derive robust stability of MPC by minimization of linear quadratic optimization problems subject to polytopic uncertainty models and linear matrix inequality (LMI) constraints, which was firstly proposed in [5]. From this formulation, a broad class of model uncertainty descriptions can be addressed with guaranteed closed-loop robust stability of MPC.

Since the Lyapunov theory was introduced as an efficient stability analysis tool of systems governed by ordinary differential equations, the notion of set invariant was used in many problems concerning the analysis and control of dynamic systems. An important motivation, leading to introduce invariant sets, was the need to analyze the effect of uncertain systems. An invariant set is a region of the state space such the trajectory generated by the dynamical system remains confined in the set if the initial condition lies within it [6]. Robust controlled invariant set is particularly relevant since it can be used in the context of constrained uncertain systems stability [7].

In recent years, in the theory of control, regardless of a particular area, there have been numerical solutions are extensive. That is, a problem is usually considered as solved whenever it can be written as a (constrained) optimization problem. The difficulty in solving such a problem is greatly influenced by the way the constraint set is defined. In this context, several families of sets vie for influence [8].

Historically, ellipsoidal sets [9] were a useful choice of invariant sets due to their simple definition.

*Corresponding Author: Laboratory of Advanced Systems (L.A.S),Polytechnic School of Tunisia,+21641594777 & hamdi.walid987@gmail.com

https://dx.doi.org/10.25046/aj030134

Then, the problem becomes to design the invariant ellipsoids off-line [10]. Recently, polyhedral sets [11], became widespread due to their representation flexibility and reliable numerical algorithmes. Angeli [12] proposed an ellipsoidal off-line MPC scheme for uncertain polytopic systems. In [13] the authors proposed an off-line robust constrained MPC algorithm by choosing a sequence of states.

However, polyhedral sets become numerically unstable for higher dimensions and certain operations scale badly with respect to the complexity of the set in question. Zonotopic sets, a subclass of polyhedral sets [11], have started to gain attention. Their symmetric shape, coupled with the flexibility inherited from the polyhedral class makes them an appealing choice for higher dimensions. Also, for dynamical systems, zonotopes provide an excellent compromise between accuracy and efficiency as first [14]. As a direct consequence, researchers from disparate fields started to employ them in various applications [15,16]. The greater part of this application exploits the zonotope facility in defining robust approximations.

Zonotopes are also used to rigorously estimate the states of dynamical systems as an alternative to observers that optimize with respect to the best estimate, such as Kalman filters. One of the first works that use zonotopes for state-bounding observers is [17] and bounded disturbance in [18]. Similarly to reachability analysis, this work has been extended to nonlinear systems in [19,20] and systems with uncertain parameters [21].

This paper is organised as follows, a description of the considered problem is first presented. Then, the optimal control problem for constrained uncertain systems is formulated. Its resolution procedure using zonotopic invariant sets with an interpolation step, is proposed. The efficiency of the used zonotopic invariant sets is then illustrated by two examples. Finally, the paper is concluded.

2 Problem description

The considered system is the following linear time-varying (LTV) system with polytopic uncertainty:

$$\begin{aligned} x(k+1) &= A(k)x(k) + Bu(k) \\ y(k) &= Cx(k) \end{aligned} \tag{1}$$

where $x(k)$ is the state of the plant, $u(k)$ is the control input and $y(k)$ is the plant output. We assume that:

$$\begin{aligned} &[A(k), B(k)] \in \Omega, \\ &\Omega = conv\{[A_1, B_1], [A_2, B_2], ..., [A_L, B_L]\} \end{aligned} \tag{2}$$

where *conv* is the convex hull and *Omega* is a polytope, $[A_j, B_j]$ are vertices of the polytope such that:

$$\left[A_j, B_j\right] = \sum_{j=1}^{L} \lambda_j \left[A_j, B_j\right], \ \sum_{j=1}^{L} \lambda_j = 1, \ 0 \le \lambda_j \le 1, \tag{3}$$

The aim is this research is to find a state-feedback control law:

$$u(k+i/k) = Kx(k+i) \tag{4}$$

that stabilizes (1) with the following performance cost:

$$\begin{aligned} &\min_{u(k+i/k)} \max_{[A(k+i),B(k+i)]\in\Omega, i\ge 0} J_\infty(k) \\ &J_\infty(k) = \sum_{i=0}^{\infty} \begin{bmatrix} x(k+i/k) \\ u(k+i/k) \end{bmatrix}^T \begin{bmatrix} \Theta & 0 \\ 0 & R \end{bmatrix} \begin{bmatrix} x(k+i/k) \\ u(k+i/k) \end{bmatrix} \end{aligned} \tag{5}$$

subject to :

$$|u_h(k+1/k)| \le u_{h,\max}, h = 1, 2, ..., n_u \tag{6}$$

$$\left|y_r(k+1/k)\right| \le y_{r,\max}, r = 1, 2, ..., n_y \tag{7}$$

where $\Theta > 0$ and $R > 0$ are symmetric weighting matrices.

In [13] the authors describe the concept of an asymptotically stable invariant ellipsoid to develop a robust constrained MPC algorithm. This algorithm gives a sequence of explicit control laws corresponding to a sequence of asymptotically stable invariant ellipsoids constructed off-line one within another in state space. They solved, at each time step, the robust constrained MPC problem using Linear Matrix Inequalities (LMI). The obtained result is considered conservative due to invariant ellipsoids which are an approximation of the real invariant sets.

In [5] the authors describe polyhedral invariant sets with an off-line robust algorithm to stabilize uncertain systems. They are calculated off-line a sequence of state feedback control laws corresponding to a sequence of polyhedral invariant sets. At each sampling time, the smallest polyhedral invariant set that the currently measured state can be embedded is determined. The corresponding state feedback control law is then implemented to the process.

We intend to use this algorithms with zonotopic representation of the invariant sets followed by an interpolation step to get less conservative results.

3 Robust MPC Algorithm

In this section, an interpolation-based robust MPC algorithm for uncertain polytopic discrete-time systems using zonotopic invariant sets is presented. The nested zonotopic invariant sets and feedback gains are pre-computed off-line in first step, in order to reduce the on-line computational burdens. In second step, the real-time control law is calculated by linear interpolation between the feedback gains corresponding to the zonotopic invariant sets previously generated. The optimization problem solved at each time step is based on optimization of linear performance index and only a computationally low-demanding optimization problem is required to be solved on-line.

Definition 1: (Invariant sets)

An invariant set is a subset of the state space $\Omega \in R^n$,

such that for all $x_0 \in \Omega$, and all admissible input function $u : R \rightarrow U$, the solution to system (1) with $x(0) = x_0$ satisfies $x(t) \in \Omega$ for all $t \geq 0$.

Intuitively, the system remains trapped in the invariant for all future times [22].

One of the advantages of invariant sets, compared with iterative methods, is that they cover unbounded time horizon, without any extra cost. A second one is that they that have in general a compact representation. For example, an invariant ellipsoid is represented by a single nn matrix. Whereas, iterative methods produce a large number of sets, often with growing complexity. Each of these sets has to be taken into account in order to enclose all reachable states.

3.1 Off-line Steps

Step 1: Choose a state sequence $x_i, i \in \{1,2,...,N\}$and solve the following problem to obtain corresponding state feedback gains:

$$K_i = Y_i Q_i^{-1} \tag{8}$$

The states x_i must be chosen such that the distance between x_{i+1} and the origin is less than the distance between x_i and the origin. Matrices Y_i and Q_i^{-1}, for all $i = 1,2,...,N$ are solutions of the following problem:

$$\min_{\gamma_i, Q_i, Y_i} \gamma_i \tag{9}$$

subject to:

$$\begin{bmatrix} 1 & x_i^T \\ x_i & Q_i \end{bmatrix} \geq 0, \tag{10}$$

$$\begin{bmatrix} Q_i & Q_i A_j^T + Y^T B_j^T & Q_i \Theta^{1/2} & Y_i^T R^{1/2} \\ A_j Q_i + B_j Y_i & Q_i & 0 & 0 \\ \Theta^{1/2} Q_i & 0 & \gamma_i I & 0 \\ R^{1/2} Y_i & 0 & 0 & \gamma_i I \end{bmatrix} \geq 0 \tag{11}$$

$\forall j = 1,2,\ldots,L$

$$\begin{bmatrix} X & Y_i \\ Y_i^T & Q_i \end{bmatrix} \geq 0,\ X_{hh} \leq u_{h,\max}^2,\ h = 1,2,\ldots,n_u \tag{12}$$

$$\begin{bmatrix} S & C(A_j Q_i + B_j Y_i) \\ (A_j Q_i + B_j Y_i)^T C^T & Q_i \end{bmatrix} \geq 0,\ S_{rr} \leq y_{r,\max}^2, \tag{13}$$
$r = 1,2,\ldots,n_y,\ \forall j = 1,2,\ldots,L,$

Step 2: Given the state feedback gains:

$$K_i = Y_i Q_i^{-1},\ i \in \{1,2,\ldots,N\} \tag{14}$$

from step 1. For each K_i, the corresponding polyhedral invariant sets defined by:

$$S_i = \{x_i / M_i x_i \leq d_i\} \tag{15}$$

are constructed by the following :

Step 2.1: Set $M_i = \left[C^T, -C^T, K_i^T, -K_i^T\right]^T, d_i = \left[y_{\max}^T, y_{\min}^T, u_{\max}^T, u_{\min}^T\right]^T$ and $m = 1$.

Step 2.2: Select row m from (M_i, d_i) and check whether $M_{i,m}(A_j + B_j K_i)x \leq d_{i,m}$ is redundant with respect to the constraints defined by (M_i, d_i) by solving the problem:

$$\max_x\ W_{i,m,j} \tag{16}$$

subject to

$$W_{i,m,j} = M_{i,m}(A_j + B_j K_i)x - d_{i,m}, \quad M_i x \leq d_i \tag{17}$$

Step 2.3: Let $m = m + 1$ and return to Step 2.2. If m is strictly larger than the number of rows in (M_i, d_i) then terminate.

3.2 On-line Step using polyhedral sets

3.2.1 Without interpolation

At each sampling time, determine the smallest polyhedral invariant set $S_i = \{x_i / M_i x_i \leq d_i\}$ where $i = 1,2,...,N-1$.
containing the measured states and implement the corresponding state feedback control law $u(k/k) = K_i x(k/k)$ to the process.

3.2.2 With 3-points interpolation

At each sampling time, if the measured state lies between S_i, S_{i+1} and S_{i+2}, $i = 1,2,...,N-1$ implement the interpolated gain obtained by :

$$\mathrm{K} = \alpha_1 K_{i-2} + \alpha_2 K_{i-1} + \alpha_3 K_i \tag{18}$$

where $0 < \alpha_i < 1$,for all $i = 1,2,3$ and $\sum_{i=1}^{3} \alpha_i = 1$.

3.3 On-line Step using zonotopic sets

Zonotopes are convex polytopes that are centrally symmetric. Equivalently, a zonotope is a Minkowski sum of a finite set of line segments. A polytope is a zonotope if it can also be represented by so-called generators (G-representation).

Definition 2: (G-representation of a zonotope)
Given a vector $c \in R^n$ and a set of vectors of R^n,$G = \{g_1,...,g_m\}, m \geq n$, a zonotope Z of order m is defined as following:

$$Z = \left\{ x \in R^n, x = c + \sum_{i=1}^{p} \gamma_i . g_i ; -1 \leq \gamma_i \leq 1 \right\} \tag{19}$$

The vector c is called the center of the zonotope Z. The vectors $g_1,...,g_m$ are called generators of Z.
The order of zonotope is defined by the number of its generators (m in this case). In the case of $m < n$, its called degenerated zonotope.

This definition is equivalent with the definition of zonotopes by the Minskowski sum of a finite number of line segments defined by $g_i B^1$. $Z = (c; g_1, g_2, ..., g_m) = c \oplus g_1 B^1 \oplus ... \oplus g_m B^1$ Where B^n is a unitary box in R^n, is a box composed by n unitary intervals. And $\oplus$ is the Minkowski sum.

Definition 3: (Minkowski sum)

The Minkowski sum of two sets X and Y is defined by $X \oplus Y = \{x + y : x \in X, y \in Y\}$.
Definition 4: (Unitary interval)
The unitary interval is defined by $B^n = [-1, 1]$.
Definition 5: (Box)
A box is an interval vector. An interval hull of a set $Z \subseteq R^n$, denoted by $\diamond Z$ is a box that satisfies $Z \subseteq \diamond Z$ Given a box $\diamond Z = ([a_1, b_1], ..., [a_n, b_n])^T$. $mid(\diamond Z)$, denotes its center and $diam(\diamond Z) = (b_1 - a_1, ..., b_n - a_n)^T$
Definition 6: (Unitary box)
A unitary box, denoted by B^n is a box compound by n unitary intervals.
Definition 7: (V-representation of a polytope)
Given r vertices $v_i \in R^n$, $P = conv\{v_1, ..., v_r\}$ is a convex polytope, where conv is the convex hull operator.
To obtain zonotopic sets from polyhedral ones, we have to perform the following three steps:
Step 1: Compute the vertices $v_i \in R^n$ (V-representation) of all N polytopes $S_i, i = 1, , N$.
Step 2: Obtain the minimum and maximum values of each polytope i:

$$\begin{aligned} m_{\min} &= \min(V_i^1, ..., V_i^v), \\ m_{\max} &= \max(V_i^1, ..., V_i^v). \end{aligned} \tag{20}$$

where V_i^j is the ith component of v^j and r is the number of the vertices of each polytope.
Step 3: Compute a G-representation of the n-dimensional interval $[m_{\min}, m_{\max}]$:

$$[m_{\min}, m_{\max}] = \left\{ x = c + \sum\nolimits_{i=1}^{n} \gamma_i . g_i, -1 \le \gamma_i \le 1 \right\}, \tag{21}$$

where :

$$c = 0.5(m_{\min} + m_{\max}), \tag{22}$$

$$g_i^{(i)} = \begin{cases} 0.5(m_{\max} - m_{\min}), & if\ i = j \\ 0, & otherwise \end{cases} \tag{23}$$

3.3.1 Without interpolation

At each sampling time, determine the smallest invariant zonotope
$Z = \left\{ x | x = c + \sum\limits_{i=1}^{p} \gamma_i . g_i, -1 \le \gamma_i \le 1 \right\}, i = 1, 2, ..., N-1$
containing the measured states and implement the corresponding state feedback control law $u(k/k) = K_i x(k/k)$ to the process.

3.3.2 With 2-points interpolation

At each sampling time, if the measured state lies between Z_i and Z_{i-1}, implement the interpolated gain obtained by :

$$K = \alpha K_i + (1 - \alpha) K_{i+1} \tag{24}$$

where $0 < \alpha_i < 1$, for all $i = 1, 2$, and $\sum\limits_{i=1}^{2} \alpha_i = 1$.

3.3.3 with 3-points interpolation

At each sampling time, if the measured state lies between Z_i,Z_{i-1} and Z_{i-2}, implement the interpolated gain obtained by:

$$K = \alpha_1 K_{i-2} + \alpha_2 K_{i-1} + \alpha_3 K_i \tag{25}$$

where $0 < \alpha_i < 1$,for all $i = 1, 2, 3$, and $\sum\limits_{i=1}^{3} \alpha_i = 1$.

4 Application

In this section, we are going to present two examples allowing to implement the proposed approach. For both examples, the software Yalmip toolbox [23] in the MATLAB environment was used to compute the solution of the LMI minimization problem.

4.1 Example 1

Lets consider an uncertain non-isothermal CSTR [5] where the exothermic reaction AB takes place. The reaction is irreversible and the rate of reaction is first order with respect to component A. A cooling coil is used to remove heat that is released in the exothermic reaction. The uncertain parameters are: the reaction rate constant k_0 and the heat of reaction ΔH_{rxn}. The linearized model based on the component balance and the energy balance is given by the following state equations:

$$\begin{cases} x(t+1) = A(t)x(t) + B(t)u(t) \\ y(t) = Cx(t) \end{cases} \tag{26}$$

where $x = \begin{bmatrix} C_A \\ T \end{bmatrix}$ is the state vector $x(t)$ and $u = \begin{bmatrix} C_{A,F} \\ F_C \end{bmatrix}$ is the control input vector $u(t)$. Matrices are defined by:

$$A = \begin{pmatrix} -\frac{F}{V} - k_0 e^{-E/RT_s} & -\frac{E}{RT_s^2} k_0 e^{-E/RT_s} C_{As} \\ \frac{-\Delta H_{rxn} k_0 e^{-E/RT_s}}{\rho C_p} & -\frac{F}{V} - \frac{UA}{V \rho C_p} \\ & -\Delta H_{rxn} \frac{E}{\rho C_p RT_s^2} k_0 e^{-E/RT_s} C_{As} \end{pmatrix} \tag{27}$$

$$B = \begin{bmatrix} \frac{F}{V} & 0 \\ 0 & -2.098 \times 10^5 \frac{T_s - 365}{V \rho C_p} \end{bmatrix}, \tag{28}$$

$$C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{29}$$

Where C_A is the concentration of A in the reactor, $C_{(A,F)}$ is the feed concentration of A, T is the reactor temperature, and F_C is the coolant flow. The operating parameters are: $F = 1m^3/\min$, $V = 1m^3$, $k_0 = 10^9 - 10^{10}\min^{-1}$, $\frac{E}{R} = 8330.1K$, $-\Delta H_{r\times n} = 10^7 - 10^8 cal/kmol$, $\rho = 10^6 g/m^3, UA = 5.34 \times 10^6 cal/(k\min)\ and\ C_p = 1cal/(gk)$. Let $\overline{C_A} = C_A - C_{A,eq}$, $\overline{T_A} = T - T_{eq}$, $\overline{C_{A,F}} = C_{A,F} - C_{A,F,eq}$ and $\overline{F_C} = F_C - F_{C,eq}$ where the subscript eq is used to denote the corresponding variable at equilibrium condition. By discretization, using a sampling time of 0.15 min, the

discrete-time model with $\begin{bmatrix} \overline{C}_A(k) \\ \overline{T}(k) \end{bmatrix}$ as a state vector and $\begin{bmatrix} \overline{C}_{A,F}(k) \\ \overline{F}_C(k) \end{bmatrix}$ as a control vector, is given as follows:

$$\begin{cases} x(k+1) = A(k)x(k) + B(k)u(k) \\ y(k) = Cx(k) \end{cases} \quad (30)$$

where:

$$A = \begin{bmatrix} 0.85 - 0.0986\alpha(k) & -0.0014\alpha(k) \\ 0.9864\alpha(k)\beta(k) & 0.0487 + 0.01403\alpha(k)\beta(k) \end{bmatrix}$$
$$B = \begin{bmatrix} 0.15 & 0 \\ 0 & -0.912 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (31)$$

where: $1 \leq \alpha(k) = k_0/10^9 \leq 10$ and $1 \leq \beta(k) = -\Delta H_{rxn}/10^7 \leq 10$.

The two parameters (k) and (k) are independent of each other. Then, we consider the following polytopic uncertain model with four vertices:

$$\Omega = Co\left\{ \begin{bmatrix} 0.751 & -0.0014 \\ 0.986 & 0.063 \end{bmatrix}, \begin{bmatrix} 0.751 & -0.0014 \\ 9.864 & 0.189 \end{bmatrix}, \begin{bmatrix} 0.751 & -0.0014 \\ 0.986 & 0.063 \end{bmatrix}, \begin{bmatrix} 0.751 & -0.0014 \\ 9.864 & 0.189 \end{bmatrix}, \right\} \quad (32)$$

The objective is to regulate the concentration $\overline{C}_A$ and the reactor temperature $\overline{T}$ to the origin by manipulating $\overline{C}_{A,F}$ and $\overline{F}_C$, respectively. These variables are constrained by: $\left|\overline{C}_{A,F}\right| \leq 0.5\, kmol/m^3$, and $\left|\overline{F}_C\right| \leq 1.5\, m^3/min$.

The cost function is given by (5) with $\Theta = I$ and $R = 0.1I$.

Lets choose a sequence of states:

$$x_i = \left\{ \begin{matrix} (0.0525, 0.0525), (0.0475, 0.0475), \\ (0.0425, 0.0425), (0.0375, 0.0375), \\ (0.0325, 0.0325), (0.0275, 0.0275) \end{matrix} \right\} \quad (33)$$

This sequence is used to calculate six off-line feedback gains K_i, $i = 1, 2, ..., 6$. The regulated output (the concentration of A and the reactor temperature), when $\alpha(k)$ and $\beta(k)$ are randomly time-varying between $10^9 \leq \alpha(k) = 10^{10}$ and $10^7 \leq \beta(k) = \Delta H_{r\times n} \leq 10^8$.

The obtained zonotopes are defined by:

$$c_i = \left\{ \; 2.98, 3.17, -1.31, 1.31, -3.17, -2.98 \; \right\}, \quad (34)$$

Where c_i is the center of the zonotope Z_i, $i = 1, 2, .., 6$.

The generators matrices are defined by:

$$g_i = \begin{pmatrix} 3.07 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.26 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.29 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.29 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.26 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.07 \end{pmatrix} \quad (35)$$

for all $i = 1, 2, .., 6$.

The regulated outputs are shown respectively in Figure 1 and Figure 2. It is seen that the considered zonotopic sets give less conservative results and better system performance as compared to the approach using polyhedral ones.

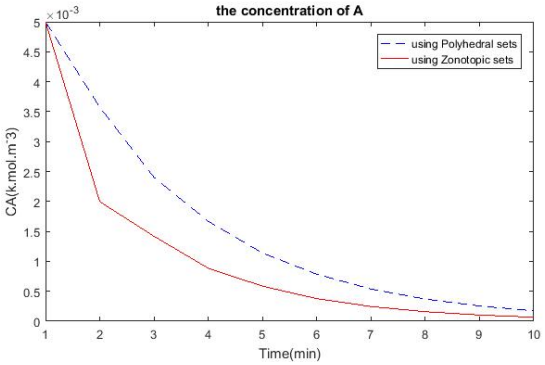


Figure 1: The concentration of A in the reactor of the

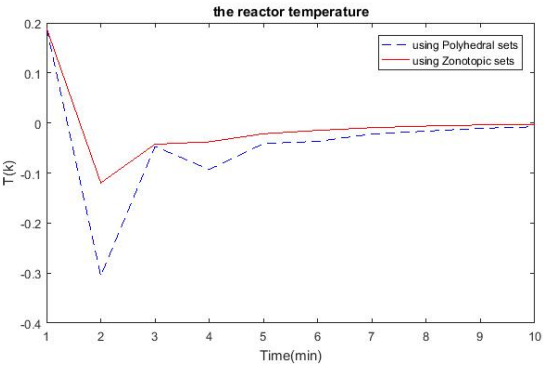


Figure 2: The reactor temperature of the regulated output

In Figure 3 and Figure 4 respectively, it is seen that the considered interpolation using three zonotopic sets, give less conservative results as compared to the approach with interpolation of polyhedral sets.

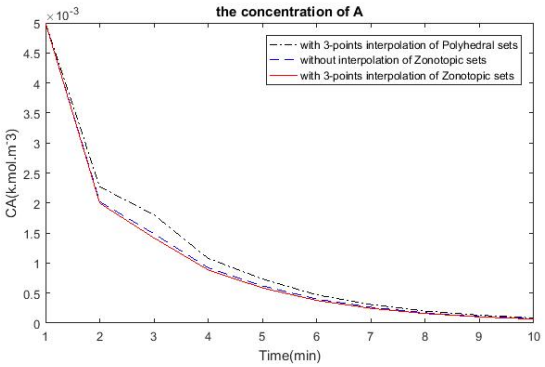


Figure 3: The concentration of A in the reactor of the regulated output

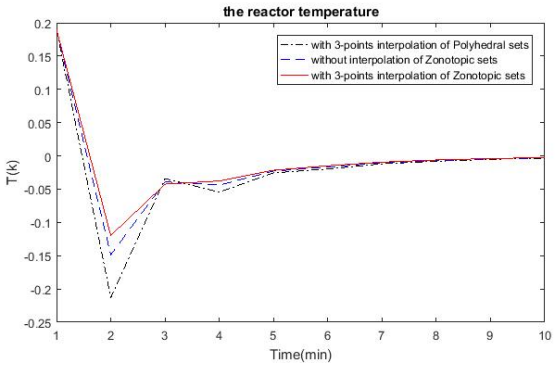


Figure 4: The reactor temperature of the regulated output

4.2 Example 2

We consider the angular positioning system [4]. It consists of an electric motor driving a rotating antenna so that it always points in the direction of a moving object. The motion of the antenna can be described by the following discrete time-equation:

$$\begin{cases} \begin{bmatrix} \theta(k+1) \\ \dot{\theta}(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 0.1 \\ 0 & 1-0.1\alpha(k) \end{bmatrix} \begin{bmatrix} \theta(k) \\ \dot{\theta}(k) \end{bmatrix} \\ \qquad\qquad + \begin{bmatrix} 0 \\ 0.0787 \end{bmatrix} u(k) \\ y(k) = [1\ 0] \begin{bmatrix} \theta(k) \\ \dot{\theta}(k) \end{bmatrix} \end{cases} \tag{36}$$

where $\theta(k)$ is the angular position of the antenna, $\dot{\theta}(k)$ is the angular velocity and $u(k)$ is the input voltage of the motor. It is assumed that the uncertain parameter is arbitrarily time-varying : $0.1\alpha(k)10$.

Let $\overline{\theta} = \theta - \theta_{eq}$, $\overline{\dot{\theta}} = \dot{\theta} - \underset{eq}{\dot{\theta}}$ and $\overline{u} = u - u_{eq}$ where the subscript eq denotes the corresponding variable at equilibrium condition. The obtained system can be written as follows:

$$\begin{cases} \begin{bmatrix} \overline{\theta}(k+1) \\ \overline{\dot{\theta}}(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 0.1 \\ 0 & 1-0.1\alpha(k) \end{bmatrix} \begin{bmatrix} \overline{\theta}(k) \\ \overline{\dot{\theta}}(k) \end{bmatrix} \\ \qquad\qquad + \begin{bmatrix} 0 \\ 0.0787 \end{bmatrix} u(k) \\ y(k) = [1\ 0] \begin{bmatrix} \overline{\theta}(k) \\ \overline{\dot{\theta}}(k) \end{bmatrix} \end{cases} \tag{37}$$

The system (36) has the following polytopic structure:

$$A(k) \in conv \left\{ \begin{bmatrix} 1 & 0.1 \\ 0 & 0.9 \end{bmatrix}, \begin{bmatrix} 1 & 0.1 \\ 0 & 0 \end{bmatrix} \right\} \tag{38}$$

The input constraint is:

$$\overline{|\overline{u}(k)|} \leq 2\, volts \tag{39}$$

The weighting matrices Θ and R are given by:

$$\Theta = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \ and \ \ R = 0.00002\, I \tag{40}$$

Lets choose the following sequence of seven states:

$$x_i = \left\{ \begin{array}{l} (0.35, 0.35), (0.3, 0.3), \\ (0.25, 0.25), (0.02, 0.02), \\ (0.15, 0.15), (0.1, 0.1), (0.05, 0.05) \end{array} \right\} \tag{41}$$

This sequence is used to calculate seven state feedback gains K_i corresponding to seven polyhedral invariant sets. The obtained zonotopes are defined by their centers:

$$c_i = \left\{ \begin{array}{l} 1.52, -0.08, -0.21, 0.21, \\ 0.08, -1.52, 0.21 \end{array} \right\} i = 1, 2, ..., 7. \tag{42}$$

The zonotope generators are given by:

$$g_i = \begin{pmatrix} 0.82 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.87 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.70 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.38 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.59 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.87 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.47 \end{pmatrix} \tag{43}$$

for all $i = 1, 2, ..., 7$.

Figure 5 and Figure 6 represent closed-loop responses of the system when $\alpha(k)$ is randomly time-varying between $0.1 \leq \alpha(k) \leq 10$.

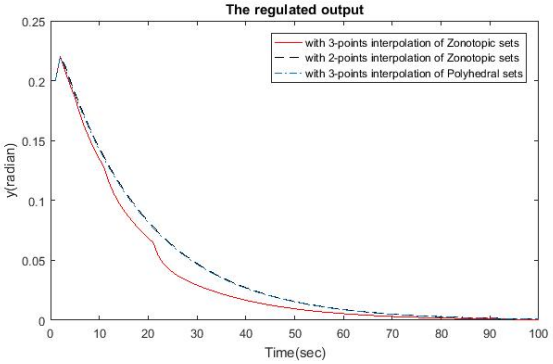


Figure 5: The regulated output

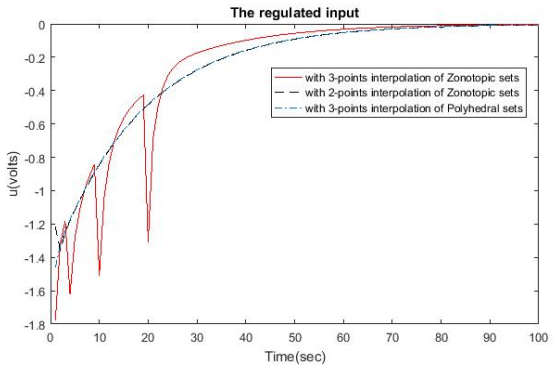


Figure 6: The control input

We can observe that by the considered approach with zonotopic sets using three points interpolation especially the one with three control gains, we obtain better control performances as compared to the approach with interpolation of polyhedral sets.

5 Conclusion

In this paper, we have presented an input feedback robust model predictive control of polytopic uncertain discrete-time systems. The proposed algorithm used an off-line optimal control optimization problems solution to determine a sequence of feedback gains. A sequence of nested zonotopic invariant sets associated with pre-computed feedback gains are constructed. At each control iteration, the smallest invariant containing the measured states is identified, and the corresponding feedback gain is implemented. In

addition, an interpolation step to the obtained control laws based on polyhedral and zonotopic invariant sets respectively was employed. The proposed approach applied on examples showed that the control performance using zonotopic invariant sets followed by an interpolation of the nested zonotopes is better than the one using polyhedral invariant sets.

References

1. S. Kheawhom and P. Bumroongsri. Interpolation-based robust constrained model predictive output feedback control, in Conference on Control and Automation. June 16-19. Palermo, Italy, 2014.

1. B. Ding, Y. Xi, M. T. Cychowski and T.O.Mahony. Improving off-line approach to robust MPC based-on nominal performance cost, Automatica, vol. 43, No. 1, pp. 158163, 2007.

2. X. Liu, S. Feng and M. Ma, Robust MPC for the constrained system with polytopic uncertainty. International Journal of Systems Science, vol. 43, No. 2, pp. 248258, 2012..

3. F. Borelli. Constrained optimal control of linear and hybrid systems, vol 290 of Lecture Notes in Control and Information Sciences, Springer, 2010.

4. M. H. Nehrir, C. Wang, Modeling and Control of Fuel Cells: Distributed Generation Applications, Wiley-IEEE Press, 2009.

5. B. Pornchai and K. Soorathep. An off-line robust MPC algorithm for uncertain polytopic discrete-time systems using polyhedral invariant sets, Journal of Process Control, vol.22, No. 5, pp. 975-983, 2012.

6. F. Blanchini and S. Miani.Set-theoretic methods in control. Systems and Control, Foundations and Applications, 2008.

7. A. Bemporad. M. Morari, V. Dua. and E. N. Pistikopoulos. The explicit linear quadratic regulator for constrained systems, Automatica, vol. 38, pp. 3-20, 2002.

8. W. Bey, Z. Kardous and N. Benhadj Braiek. Stabilization of Constrained uncertain systems by Multi-Parametric Optimization,International Journal of Automation and Control(IJAAC), vol. 10, n. 4, pp. 407-416, Inderscience Enterprises Ltd, 2016.

9. A. B. Kurzhanskii and I. Vlyi. Ellipsoidal calculus for estimation and control, Burlhauser, Boston, Massachusets, 1997.

10. A. C. Brooms, B. Kouvaritakis, and Y. I. Lee. Constrained MPC for uncertain linear systems with ellipsoidal target sets, Systems and Control Letters, vol. 44, No. 3, pp. 157166, 2011.

11. A. Matthias, S. Olaf and B. Martin. Computing reachable sets of hybrid systems using a combination of zonotopes and polytopes, Nonlinear Analysis: Hybrid Systems, vol.4, No. 2,pp. 233-249, 2010.

12. A. Casavola, D. Angeli, G. Franze and E.Mosca. An ellipsoidal off-line MPC scheme for uncertain polytopic discrete-time systems, Automatica, vol. 44, No. 12, pp. 31133119, 2008.

13. Z. Wan and M. V. Kothare, An efficient off-line formulation of robust model predictive control using linear matrix inequalities. Automatica, vol. 39, No. 5, pp. 837846, 2003.

14. A. Ingimundarson, J. M. Bravo, V. Puig, T. Alamo and P. Guerra. Robust fault detection using zonotope-based set-membership consistency, International journal of adaptive control and signal processing, vol. 23, No. 4, pp. 311330, 2008.

15. M. Althoff, O. Stursberg and M. Buss. Computing reachable sets of hybrid systems using a combination of zonotopes and polytopes, Nonlinear Analysis: Hybrid Systems, vol. 4, No. 2, pp. 233249, 2010.

16. J. Blesa, V. Puig and J. Saludes. Identification for passive robust fault detection using zonotope-based set-membership approaches, International Journal of Adaptive Control and Signal Processing, vol.25, No.9, pp. 788-812, 2011.

17. C. Combastel. A state bounding observer based on zonotopes, In Proceeding of the European Control Conference, 2003.

18. F. Stoican, S. Olaru, J. A. De Don and M. M. Seron. Zonotopic ultimate bounds for linear systems with bounded disturbances, In Proceedings of the 18th IFAC World Congress, Milano, Italy, pp. 92249229, 2011.

19. T. Alamo, J. M. Bravo and E. F. Camacho. Guaranteed state estimation by zonotopes, In Proceeding of the 42nd IEEE Conference on Decision and Control, pp. 58315836, 2003.

20. C. Combastel. A state bounding observer for uncertain nonlinear continuous-time systems based on zonotopes, In Proceeding of the 44th IEEE Conference on Decision and Control, and the European Control Conference, ECC, pp.72287234, 2005.

21. V. T. H. Le, C. Stoica, T. Alamo, E. F. Camacho and D. Dumur. Zonotopic guaranteed state estimation for uncertain systems. Automatica, vol. 49, No. 11, pp.4183424, 2013.

22. I. Ben Makhlouf, P. Hansch and S. Kowalewski. Comparison of Reachability Methods for Uncertain Linear Time-Invariant Systems, European Control Conference (ECC) July 17-19 Zrich, Switzerland, 2013.

23. J. Lofberg, Yalmip: A toolbox for modeling and optimization in matlab, in Proc. IEEE international symposium on computer aided control systems design, pp. 284289, 2004.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 297-309 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

The Use of LMS AMESim in the Fault Diagnosis of a Commercial PEM Fuel Cell System

Reem Izzeldin Salim[*,1], Hassan Noura[2], Abbas Fardoun[3]

[1]*Electrical Engineering Department, United Arab Emirates University, Al Ain, United Arab Emirates*

[2]*Faculty of Engineering, Islamic University of Lebanon, Wardanyeh, Lebanon*

[3]*Electrical Engineering Department, Al Maaref University, Beirut, Lebanon*

ARTICLE INFO

Article history:
Received: 30 October, 2017
Accepted: 18 January, 2018
Online: 30 January, 2018

Keywords:
Proton Exchange Membrane Fuel Cells
Modeling
Simulation
AMESim
Fault Diagnosis

ABSTRACT

The world's pollution rates have been increasing exponentially due to the many reckless lifestyle practices of human beings as well as their choices of contaminating power sources. Eventually, this lead to a worldwide awareness on the risks of those power sources, and in turn, a movement towards the exploration and deployment of several green technologies emerged.

Proton Exchange Membrane Fuel cells (PEMFCs) are one of those green technologies. However, in order to be able to successfully and efficiently deploy PEMFC systems, a solid fault diagnosis scheme is needed. The development of accurate model based fault diagnosis schemes has been imposing a lot of challenge and difficulty on researchers due to the high complexity of the PEMFC system. Furthermore, confidentiality issues with the manufacturer can also impose further constraints on the model development of a commercial PEMFC system. In this work, an approach to develop an accurate PEMFC system model despite the lack of crucial system information is presented through the use of Siemens LMS AMESim software. The developed model is then used to develop a fault diagnosis scheme that is able to detect and isolate five system faults.

1. Introduction

Proton Exchange Membrane Fuel Cells are complex multi-physics systems (chemical, electrical, fluidic, thermal, and mechanical phenomena are inter-acting with one another). This makes the modeling and fault diagnosis of PEMFC systems a very difficult task. Furthermore, when the modeling is based on experimental testing and experimental data, some limitations are usually faced. For example, many physical system data may be absent or difficult to obtain with the system's supplied data acquisition. Furthermore, due to warranty issues, the addition of sensors might be difficult since only limited access to the systems is allowed. Likewise, other specific parameters might be unobtainable due to manufacturer confidentiality issues.

This paper is an extension of work originally presented in the 7th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO) [1]. It presents a novel approach, in which the Siemens software, LMS AMESim 14, is used as an alternative modeling tool to model a PEMFC system when such limitations are faced. The LMS AMESim software can serve as an excellent and attractive simulation and modeling platform for different complex physical systems such as automobile and aviation systems as well as power generation systems and transmission lines. Moreover, it has outstanding simulation capabilities that makes it an excellent platform for fault simulation, and fault diagnosis studies. The graphical user interface comprehensive library it contains makes it very easy to compile a complete system model from different system components. Instead of writing all the modeling equations which could be time consuming and is prone to modeling errors, a user

[*]Reem I. Salim, Electrical Engineering Dept., UAEU, 15551 Al Ain, UAE, Tel: +971507649740, Email: reemsalim@gmail.com

can use a ready accurate model better and focus on the parameter identification step in order to find the model that matches the pursued system performance measures. Furthermore, a researcher can easily change the scripts of the components used to better suit their system's design, as well as create new libraries with more specific components.

The ElectraGen™ 3 kW PEMFC system shown in Figure 1, is an actual commercial system used in practice in many sectors, especially in telecommunication companies. It uses hydrogen gas supplied from pressurized hydrogen cylinders as the anode fuel, and atmospheric air supplied by a compressor and humidified through a built-in humidity exchanger as the cathode fuel. The ElectraGen™ system is an outdoor air cooled system, and it is only operable at ambient temperatures ranging from –40°C to 50°C. It contains a total of 38 cells and can produce up to 3 kW of unregulated DC output power and has a rated voltage of 48V.

Figure 1: The ElectraGen™ PEMFC system and the 3 kW Load.

The module contains the ElectraGen™ 3 kW Fuel Cell stack with integrated microprocessor controller and safety features, a hydrogen pressure gauge which gives an indication of fuel level and the 3kW load which consists of 30 lamps, 100W each (see Figure 1). The system is connected through data acquisition to a GUI based on LabVIEW to monitor and log the different system variables (stack current, stack voltage, external voltage, individual cell voltages, cabinet temperature, cathode air temperature, coolant temperature, exhaust temperature, and hydrogen pressure). The ElectraGen™ system is installed outdoor and the experimental data sets collected from the system were taken at extreme environmental conditions in the summer at noon with the ambient temperature ranging from 48°C to 52°C. However, due to warranty issues, limited access to the systems was allowed and limited data was obtainable from the data acquisitions. Moreover, due to confidentiality issues with the manufacturer, several physical parameters were unobtainable such as active cell area, membrane length, volumes of cathode and anode chambers, mass of stack, etc. This made it very difficult to model the system and develop an accurate fault diagnosis scheme. Therefore LMS AMESim was used as an alternative modeling platform since all the common PEMFC modeling equations available in literature [2, 3] are already embedded in its library components.

Furthermore, it is convenient to mention here that the simulated model will not be an exact match for the ElectraGen™ system since not enough system readings were available to match the model to. However, it is the aim of this work to develop an AMESim model that is as realistic as possible by matching all the known features of the actual physical system (the ElectraGen™ system) to their equivalents in the AMESim model as best as possible. This model can then be used in the fault diagnosis study.

Section 2 presents the ElectraGen™ modeling and validation results using the Siemens LMS AMESim Software. In section 3, the simulation of five different system faults is presented, and in section 4 two residual generation techniques are evaluated and then outperforming technique is used to develop a fault diagnosis scheme in AMESim for the ElectraGen™ system. The fault diagnosis scheme is then evaluated and concluding remarks are finally given in section 5.

2. AMESim Modeling of the ElectraGen™ 3 kW System

The LMS AMESim software contains several embedded parameter identification tools including Genetic Algorithms (GA). However, in order to be able to use such tools efficiently, several software licenses are needed. To further explain this, GA is a population based mechanism that is known to be successful because it performs parallel evaluations, and when only one AMESim license is available, using GA becomes impractical. As an example, if the GA has 20 individuals in its population, and with a preset maximum number of generations of 500, this is equivalent to 10,000 runs of the software. With one system license, the software will be unable to perform parallel evaluations. Thus, if a single run of the software takes 1 minute, then the parameter identification process using GA and one system license will take 10,000 minutes (almost seven days).

As a result, matching the parameters of the AMESim model to the actual system performance of the ElectraGen™ system was done through trial and error. The LMS AMESim model developed for the ElectraGen™ PEMFC system is presented in Figure 2.

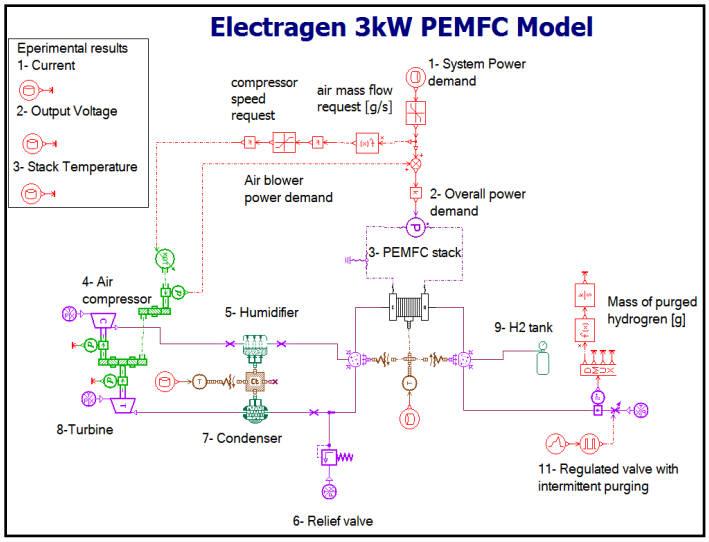


Figure 2: ElectraGen™ 3 kW system's model in AMESim.

2.1. AMESim Model Parameter Identification

The actual power demand data collected experimentally from the ElectraGen™ system was used as the input to the system in

Figure 2. After several trial and error attempts to match the stack voltage, stack current, and stack temperature values as best as possible to the actual experimental data collected from the system; the best obtained match is depicted in the following figures: Figure 3 gives the power demand input of the modeling data set, and Figures 4, 5 and 6 presents a comparison between the actual experimental current, stack voltage and stack temperature respectively to those resulting from the AMESim simulation model.

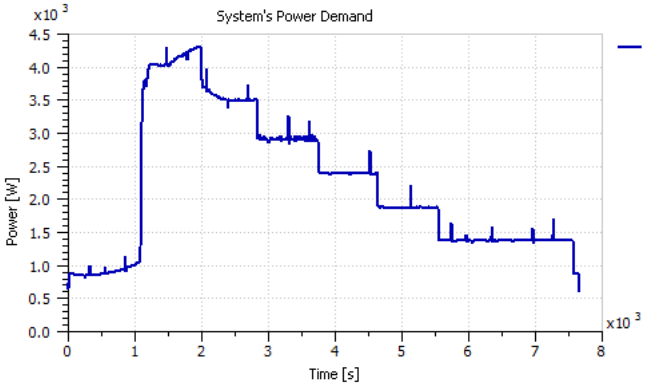


Figure 3: Power demand of the AMESim modeling data set.

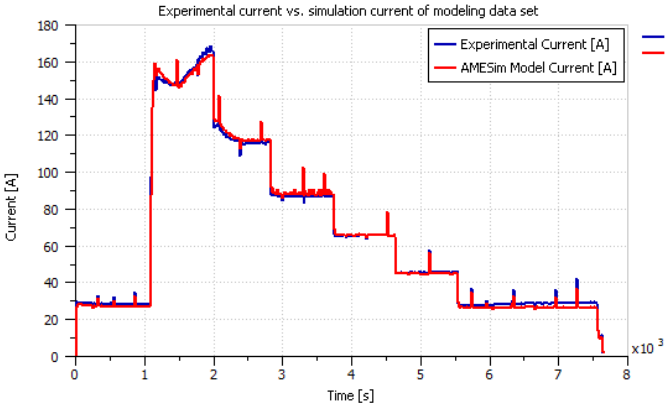


Figure 4: Experimental versus AMESim model's resulting current of the modeling data set.

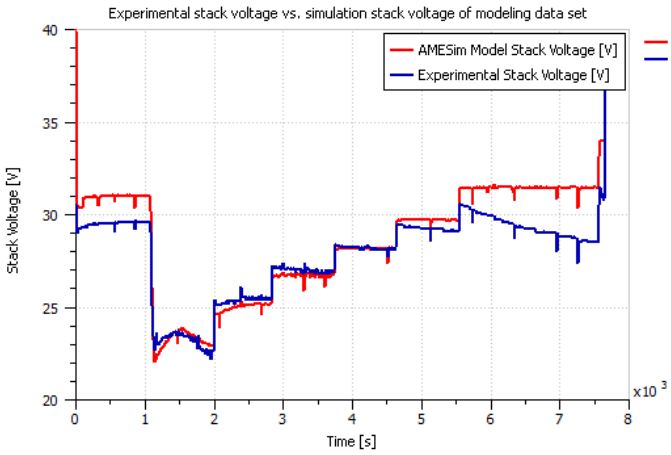


Figure 5: Experimental versus AMESim model's resulting stack voltage of the modeling data set.

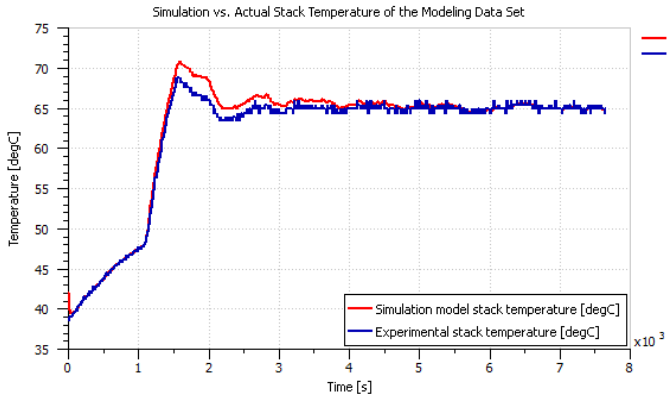


Figure 6: Experimental versus AMESim model's resulting stack temperature of the modeling data set.

2.2. AMESim Model Validation

Two other ElectraGen™ experimental data sets were used to validate the obtained AMESim model of Figure 2. Figure 7 presents the power demand of the first validation example and Figures 8 and 9 compare the actual experimental current and stack voltage to those obtained from the LMS AMESim model respectively. Similarly, the power demand of the second validation example is given in Figure 10, and a comparison between the actual and simulation current is presented in Figure 11, whereas a comparison between the actual and simulation stack voltage is presented in Figure 12.

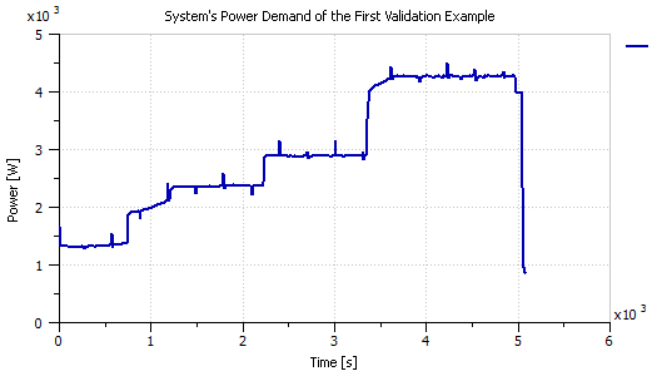


Figure 7: Power demand of the first validation example.

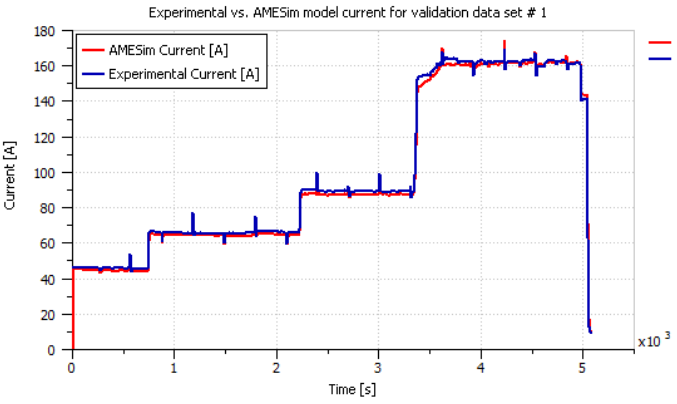


Figure 8: Actual versus simulation current of the first validation example.

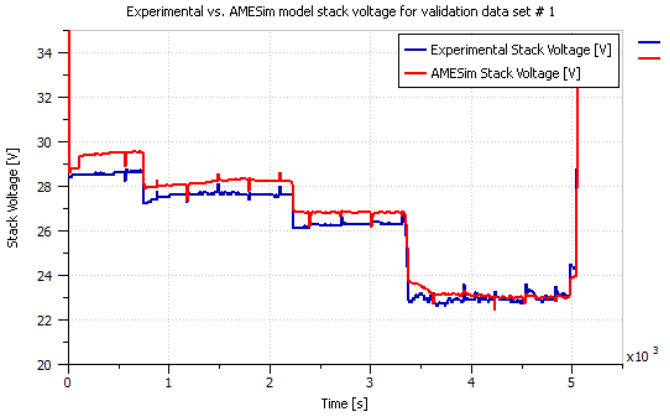


Figure 9: Actual versus simulation stack voltage of the first validation example.

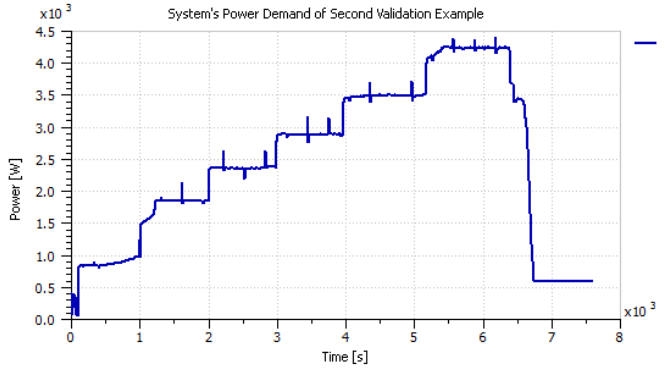


Figure 10: Power demand of the second validation example.

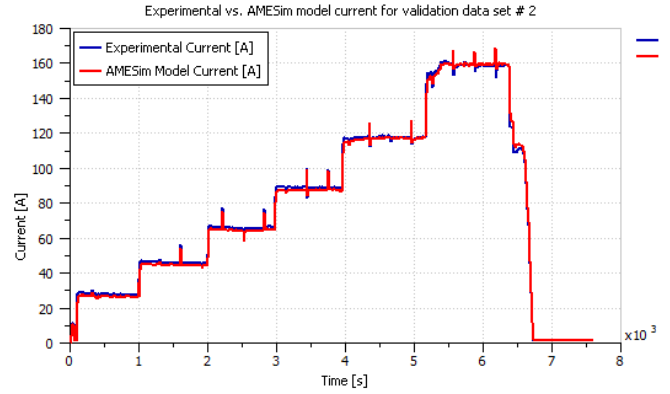


Figure 11: Actual versus simulation current of the second validation example.

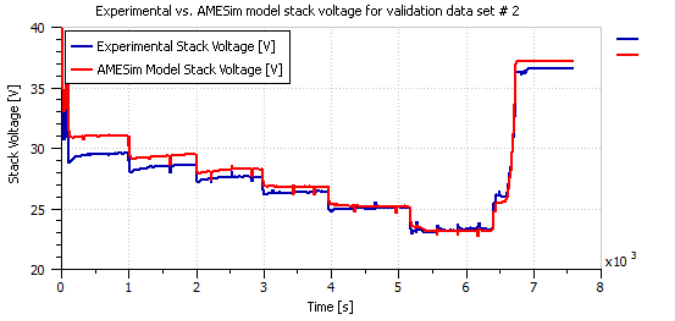


Figure 12: Actual versus simulation stack voltage of the second validation example.

Obviously, a perfect representation of the ElectraGen™ system is unobtainable through trial and error alone. Nevertheless, after validation, the AMESim obtained model proved to give a proper representation of the system's performance and was therefore used in the following fault diagnosis study.

Moreover, if resources (multiple licenses of the software) were available, a much more accurate representation of the system would have been achievable through the use of Genetic Algorithms in the parameter identification approach in AMESim, which would have in turn lead to a better fault diagnosis study.

3. Fault Simulation Using AMESim Model

Several faults were induced in the AMESim model and the system's performance measures towards those faults were recorded. The induced faults are:

1. Drying
2. Flooding
3. Air leakage
4. Hydrogen Leakage
5. Cooling System Failure

The above faults were simulated using different techniques.

3.1. Drying

Zawodzinski et al. were the first to describe the water content in the membrane by (λ) in [4] in order to estimate its state of humidity. As presented in (1), λ represents the ratio between the number of water molecules in the membrane to the number of ($SO_3^- H^+$) charge sites in the Nafion layer of the membrane [5].

$$\lambda = \frac{H_2O}{SO_3^- H^+} \quad (1)$$

The AMESim PEMFC stack module automatically calculates the water content λ in the membrane. Therefore, when simulating drying or flooding, λ would give an indication towards the membrane's state of health.

However, λ cannot be used in the fault diagnosis study because this parameter is not readily available in commercial fuel cells. Furthermore, the PEMFC undergoes drying condition when the water content in the membrane λ drops below 4 [6].

Therefore in order to simulate drying, the humidity level of the input air was set to be 0% and the target humidity level of the humidifier was dropped to 10% only. The water content in a healthy stack and that of a drying stack are compared in Figure 13. The water content of the stack that is undergoing drying is obviously well below 4. Figure 14 on the other hand depicts the difference between the polarization curve of a healthy stack and a drying stack. It is noticed that drying results in a significant voltage.

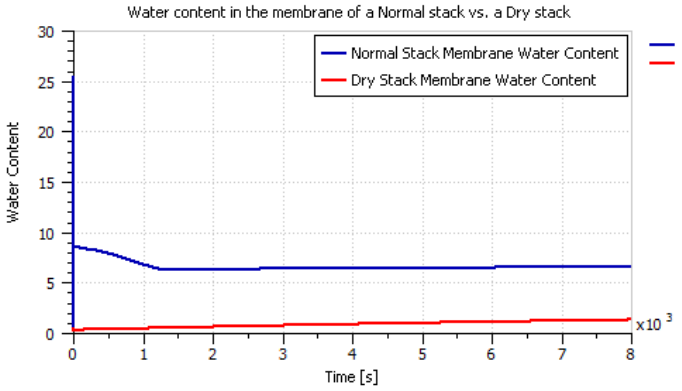


Figure 13: The water content in the membrane (λ) of a normal stack versus a dry stack.

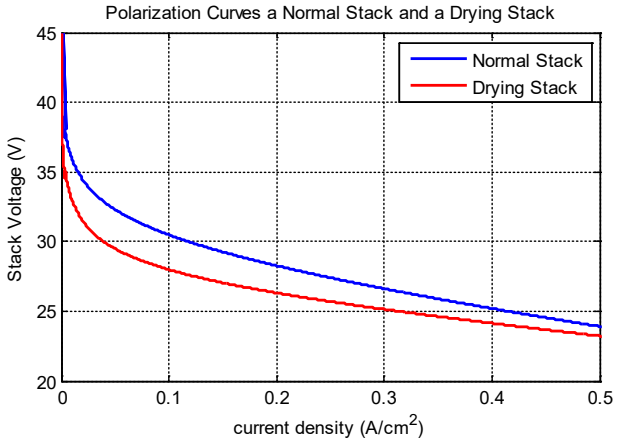


Figure 14: Normal stack's polarization curve in comparison to a dry stack's polarization curve.

3.2. Flooding

Flooding was simulated by dropping the stack temperature to 25°C while increasing the humidifier's target humidity level to 100%. Flooding affected both the voltage and current profiles of the stack as depicted in the polarization curve comparison given in Figure 15. Furthermore, Figure 16 shows the water content λ of the simulated flooded stack.

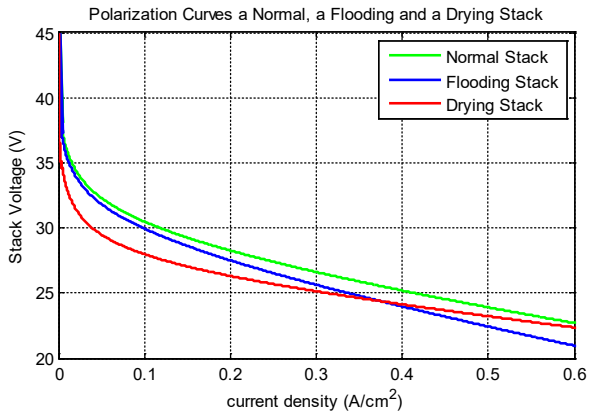


Figure 15: Comparison of pressure drop in a normal stack, a flooding stack and a drying stack.

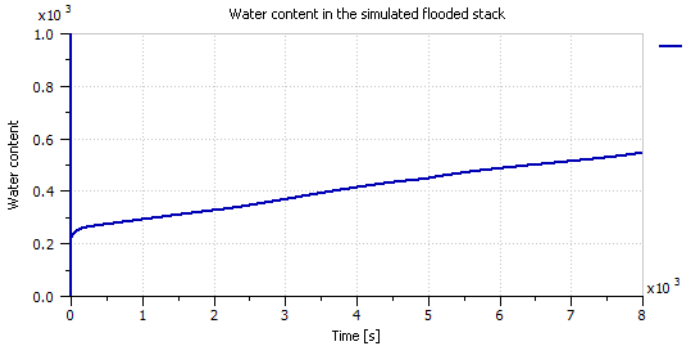


Figure 16: The water content in the membrane (λ) of the simulated flooding stack

Note that both flooding and drying result in similar effects to the stack's polarization curve. However, flooding seemed to also cause a distinctive effect on the cathode pressure drop. As the stack gets flooded with water, the pressure drop across the cathode increases. Nonetheless, in order to clearly see this effect on cathode pressure, the stack had to be fed with a step power demand profile such as that of Figure 17. The cathode pressure of both the healthy and flooded stacks with respect to the power demand input of Figure 16 are compared in Figure 17. The flooded stack showed a steeper drop in pressure when compared to the healthy stack.

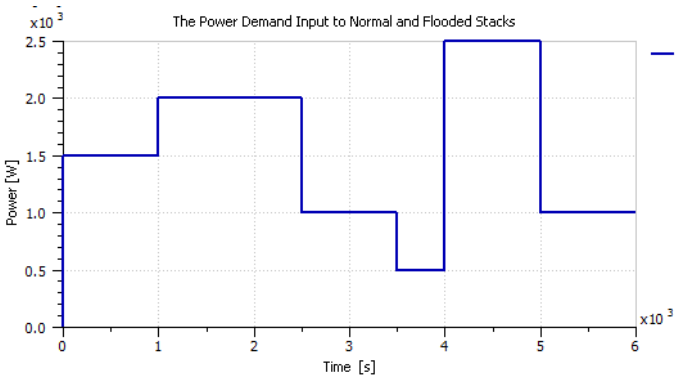


Figure 17: Power demand input to normal stack and flooding stacks.

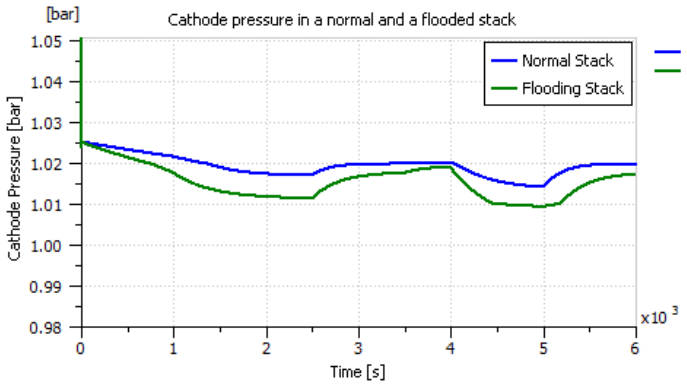


Figure 18: Comparison of pressure drop in a normal stack and a flooding stack.

3.3. Air leakage

In order to simulate air leakage, a relief valve was added right after the humidifier as depicted in Figure 19, and was set to leak air at a rate of around 10 g/s. Figure 20 shows the amount of leakage introduced in g/s. Furthermore, Figure 21 compares the input air flow rate to the PEMFC stack with and without air leakage.

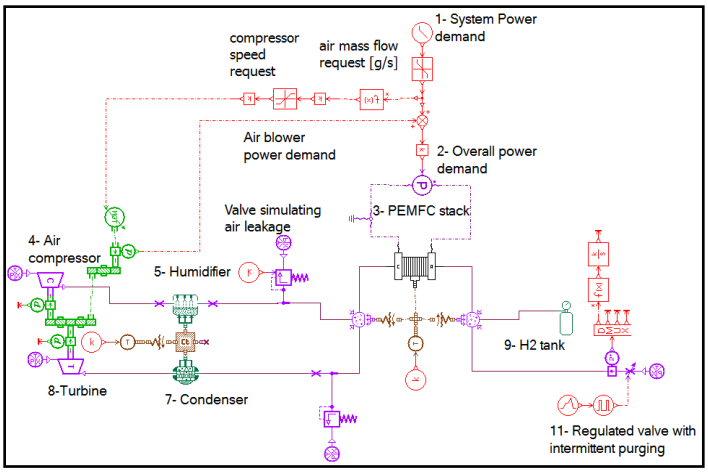


Figure 19: The relief valve added to the AMESim model to simulate air leakage.

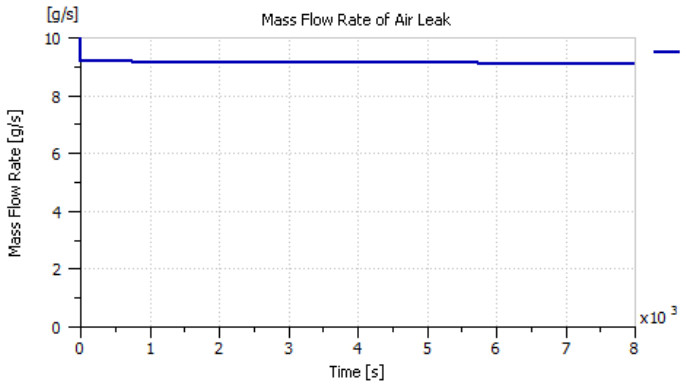


Figure 20: The flow rate of the air leakage induced to the PEMFC system.

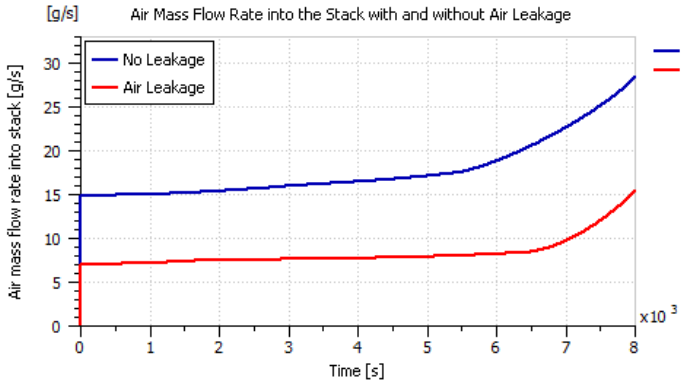


Figure 21: Input air flow rate to the stack with and without air leakage.

It can be deduced from the polarization curve comparison depicted in Figure 22 that the air leakage had no noticeable effect on the stack's current or voltage. However, it was noticed to impose a significant effect on the cathode pressure. To better see this effect, both the healthy and air leaking stacks were fed with the step power demand of Figure 17. Figure 23 shows the air leakage effect on the cathode pressure drop when compared to a non-leaking stack.

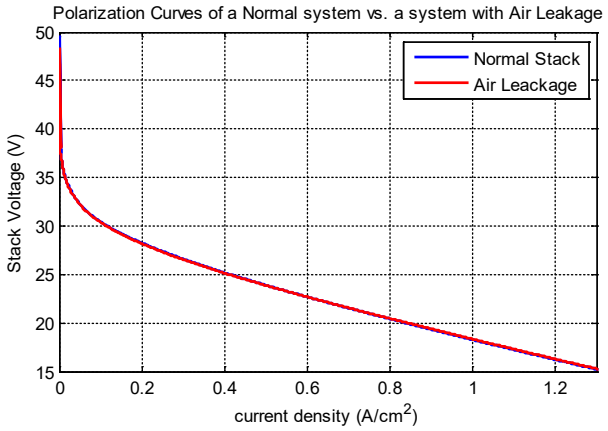


Figure 22: Polarization curves of a normal system vs. a system undergoing air leakage.

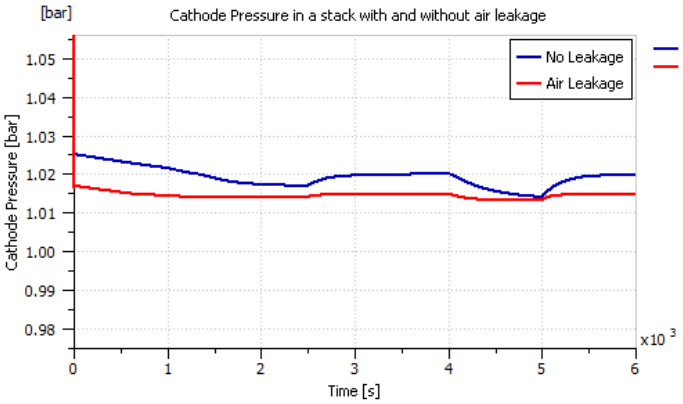


Figure 23: Pressure drop in the cathode of a normal system vs. a system undergoing air leakage.

3.4. Hydrogen leakage

Similar to air leakage, hydrogen leakage was also simulated through the addition of a relief valve right after the hydrogen canister as depicted in Figure 24 in order to leak hydrogen at a rate of 10 g/s. Figure 25 shows the amount of hydrogen leakage introduced in g/s.

Hydrogen leakage was found to result in a slight effect on stack's polarization curve as shown in the comparison of Figure 26. Furthermore, hydrogen leakage was also found to affect the anode pressure as shown in Figure 27, but had no significant effect on the cathode pressure as seen in Figure 28.

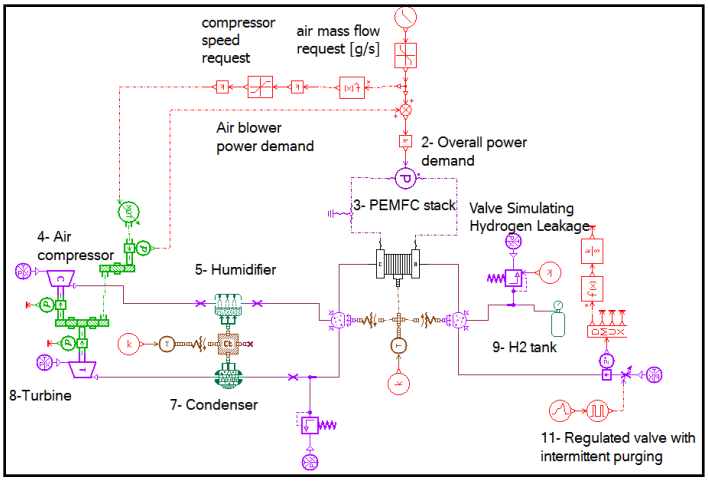


Figure 24: The relief valve added to the AMESim model to simulate Hydrogen leakage.

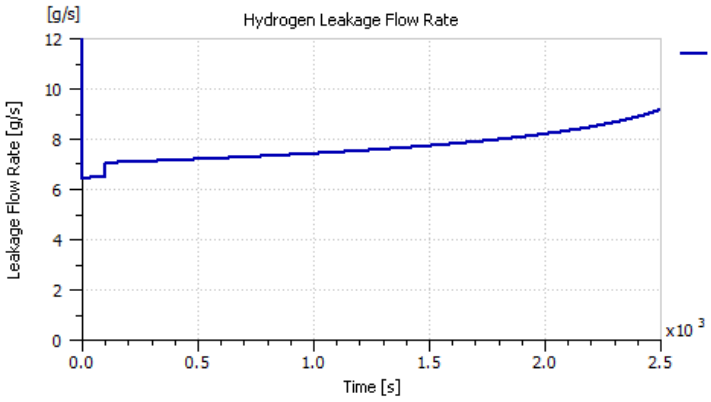


Figure 25: The flow rate of the Hydrogen leakage induced to the PEMFC system.

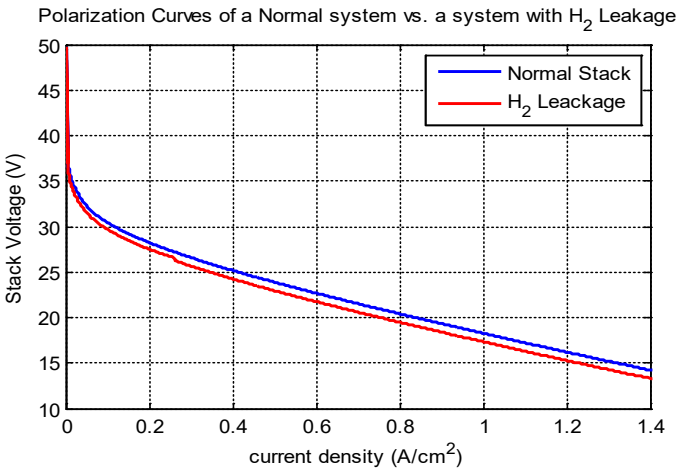


Figure 26: Polarization curves of a normal system vs. a system undergoing Hydrogen leakage.

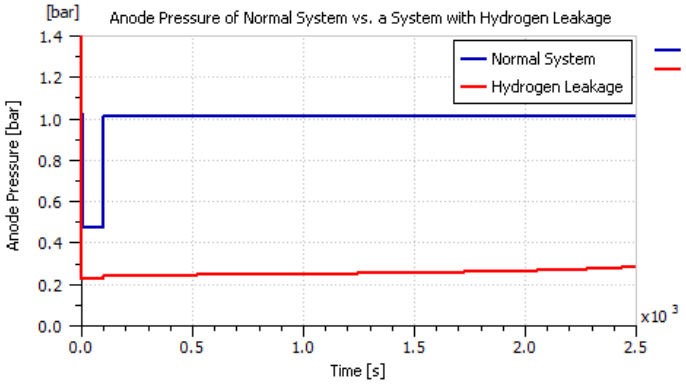


Figure 27: Anode pressure of a normal system vs. a system undergoing Hydrogen leakage.

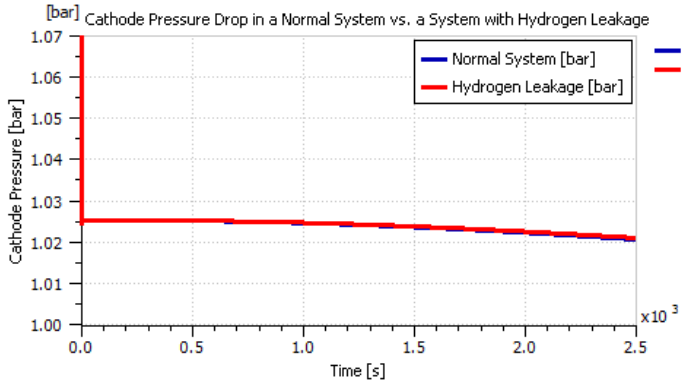


Figure 28: Pressure drop in the cathode of a normal system vs. a system undergoing Hydrogen leakage.

3.5. Cooling Failure

The stack temperature of the ElectraGen™ system should be maintained at a temperature value between 60°C to 65°C for maximum efficiency through air cooling. However, it is also convenient to mention that the stack temperature should never reach any value above 75°C in order to avoid damage of the membrane. Thus, it is important to flag a cooling system failure as soon as the stack temperature reaches 75°C or higher. Hence, the stack's temperature was increased to 75°C in order to simulate cooling failure. The polarization curve was found to undergo a significant effect with the increase in stack temperature as shown in Figure 29.

Note that the 75°C stack temperature results in a slight improvement in the system's polarization curve because the resistive components in the activation and ohmic voltage drop will decrease with the increase in stack temperature. However, at such an elevated temperature, drying of the stack will be inevitable. Furthermore, comprehensive literature review [7] revealed that operating the PEMFC at higher stack temperatures is a common stressor for almost all health degradation mechanisms. Thus, this slight improvement in the polarization curve is worthless since it will significantly shorten the PEMFCs lifespan.

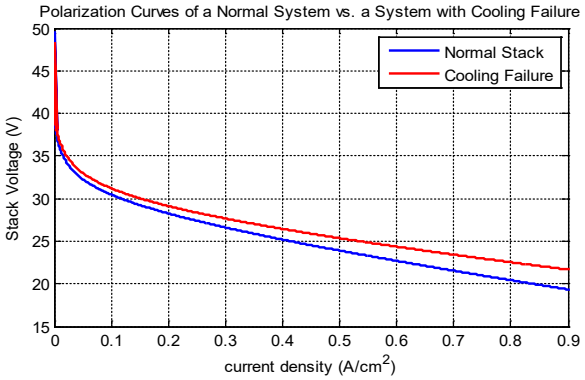


Figure 29: Polarization curve of a normal system vs. a system undergoing cooling failure.

Nonetheless, the stack temperature should be the main parameter used to detect the presence of cooling failure regardless of the effects on the voltage and current profiles. As soon the stack's temperature reading reaches 75°C is a cooling system failure fault should be flagged.

Note from Figure 29 that there is an improvement in the voltage performance of the fuel cell with the increase of stack temperature. This is expected since the rate of chemical reactions increase with the temperature causing this increase in voltage. However, it should also be noted that at this temperature value of 75°C, drying of the membrane is inevitable. Furthermore, high stack temperature is a well-known stressor for PEMFCs [8 – 11]. Thus, operating the system at such elevated stack temperatures will significantly reduce its lifespan, which makes this small voltage improvement at high stack temperature values worthless.

4. Fault Diagnosis

In this work, the model based fault diagnosis approach is based on the real-time comparison between the actual system performance and the performance predicted by the developed AMESim model. Any predicted discrepancies will be analyzed to determine the type of system fault occurring at the moment.

4.1. Residual generation

It can be concluded from the previous section that in order to detect discrepancies between the actual system and its developed model that can help in the fault detection and isolation process, five residuals (see Figure 30) should be generated based on the following five system variables: stack voltage (V_{stack}), current (I), stack temperature (T_{stack}), cathode pressure ($P_{Cathode}$) and anode pressure (P_{anode}).

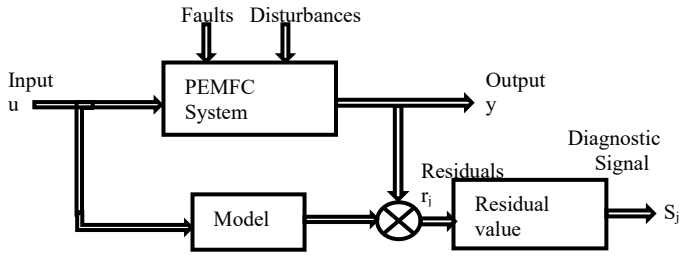


Figure 30: Residual generation diagram [12].

Several forms exist in literature for the calculation of residuals. In the simplest form given in (2), the absolute value of the residual (r_j) is compared to a relative threshold value (τ_j). The diagnostic signal (S_j) is then set to be "0" if the absolute residual is less than or equal to the threshold to indicate that no discrepancy is detected, and "1" if the absolute residual is greater than the set threshold which indicates that a discrepancy is detected [12]

$$S_j = \begin{cases} 0 & if \quad |r_j| \le \tau_j \\ 1 & if \quad |r_j| > \tau_j \end{cases} \tag{2}$$

Note that basing the fault diagnosis approach on the residuals calculated in (2) makes the approach prone to diagnosis errors. This is because the residuals of (2) are highly sensitive to instantaneous changes in the system such as the electromagnetic disturbance pulses which act on the system's measured signals as well as the occurrence of actual system faults.

Therefore, the calculation of those residuals can be slightly altered to decrease their sensitivity to disturbances. Thus, instead of calculating instantaneous residuals, they can be calculated over a moving window of time and judging the average of all residuals in that window as explained by (3), where N_r represents the number of residuals in the window [13]:

$$S_j = \begin{cases} 0 & if \quad \tilde{r}_j(N_r) = \frac{1}{N_r}\left|\sum_{n=0}^{N_r-1} r_{j,k-n}\right| \le \tau_j \\ 1 & if \quad \tilde{r}_j(N_r) = \frac{1}{N_r}\left|\sum_{n=0}^{N_r-1} r_{j,k-n}\right| > \tau_j \end{cases} \tag{3}$$

The threshold value for each residual is given in Table 1, and it was set based on the accuracy of the actual system sensors installed in the ElectraGen™ PEMFC system.

Table 1: Accuracy of actual sensors used to measure system variables

System variable	Sensor accuracy
Stack Voltage (V_{stack})	± 0.5 V
Current (I)	± 1 A
Stack Temperature (T_{stack})	± 3 °C
Cathode Pressure ($P_{cathode}$)	± 1 mbar
Anode Pressure (P_{anode})	± 1 mbar

4.2. Sensitivity assessment

Both the residual calculation techniques of (2) and (3) were implemented in LMS AMESim to study and assess their effectiveness and test their sensitivity towards electromagnetic disturbance pulses. Thus, a noisy step power demand signal with disturbance pulses was generated to help with the assessment. The system estimator however was fed with a clean power demand signal to test the diagnostic signals sensitivity toward those disturbance pulses.

The residuals that were calculated by the system are: r_1 which is based on the stack temperature (T_{stack}), r_2 which is based on the stack voltage (V_{stack}), r_3 which is based on the current (I), r_4 which is based on the anode pressure (P_{anode}) and r_5 which is based on the

cathode pressure ($P_{cathode}$). On the other hand, S_i is the respective diagnostic signal r_i.

The faults being detected by the proposed fault diagnosis approach are: (f_1 = Drying, f_2 = Flooding, f_3 = Air Leakage, f_4 = Hydrogen Leakage, and f_5 = Cooling Failure). The results of section III are summarized in Table 2 to help in the fault discrimination process, where 0 indicates that no discrepancies are detected between the predicted value and the actual measured value, and 1 indicates the positive detection of discrepancies whereas X indicates the fault exists whether the diagnostic signal detects discrepancies in the system variable or not. Note that the cooling failure that f_5, is set to be flagged as soon as the stack temperature reaches a value of 75°C (i.e. based on S_1) regardless of the effects on the stack voltage and current (i.e. S_2 and S_3).

Table 2: The effect of faults on diagnostic signals

fault	S_1	S_2	S_3	S_4	S_5
f_1	0	1	1	0	0
f_2	0	1	1	0	1
f_3	0	0	0	0	1
f_4	0	1	1	1	0
f_5	1	X	X	0	0

Figure 31 shows the fault diagnosis system built based on the instantaneous diagnostic signals calculation of (2). Figure 32 shows the noisy power demand signal fed to the actual system being diagnosed and the clean power demand signal fed to the system estimator. The diagnostic signals calculated based on the instantaneous residuals of (2) are shown in Figure 33.

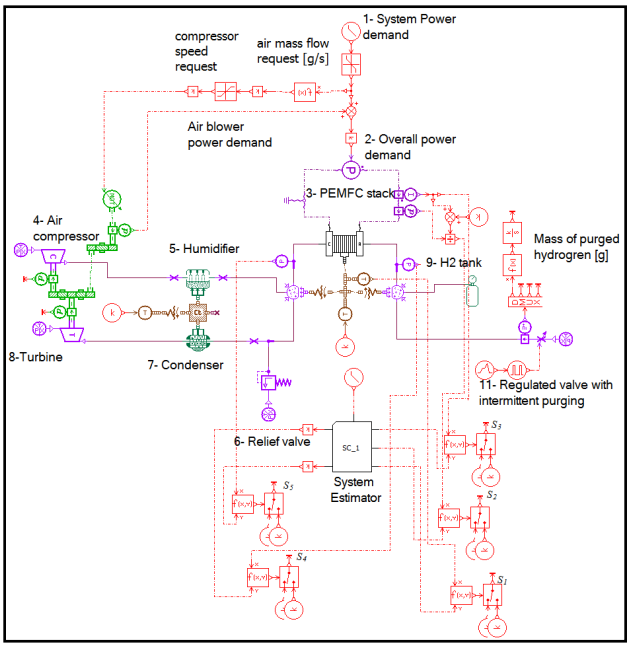


Figure 31: Fault diagnosis system based on instantaneous diagnostic signals.

It is obvious from Figure 33 that the diagnosis approach based on the instantaneous residuals in (2) is impractical and highly sensitive to electromagnetic disturbance pulses, which can therefore result in the detection of a faults when no fault actually exists. For instance, at around 150 s, both S_2 and S_3 diagnostic signals were triggered, which from Table 2 , will indicate the presence of the first fault f_1 (Drying).

Afterwards, the diagnostic signal of (3) was also implemented in LMS AMESim in order to be assessed based on three different moving windows of time (5 s, 10 s and 15 s). The same power signals of Figure 32 were used in the assessment to evaluate its sensitivity towards electromagnetic disturbance pulses. The fault diagnosis system built based on the diagnostic signals calculation of (3) is presented in Figure 34.

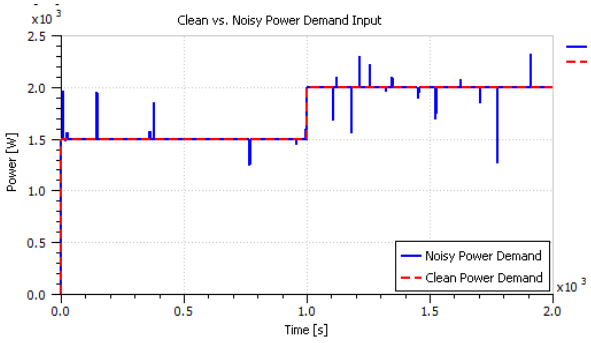


Figure 32: Clean and noisy power demand signals.

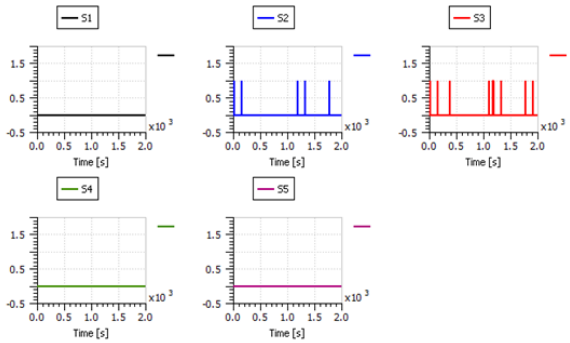


Figure 33: The effect of the pulsating noise on the five diagnostic signals.

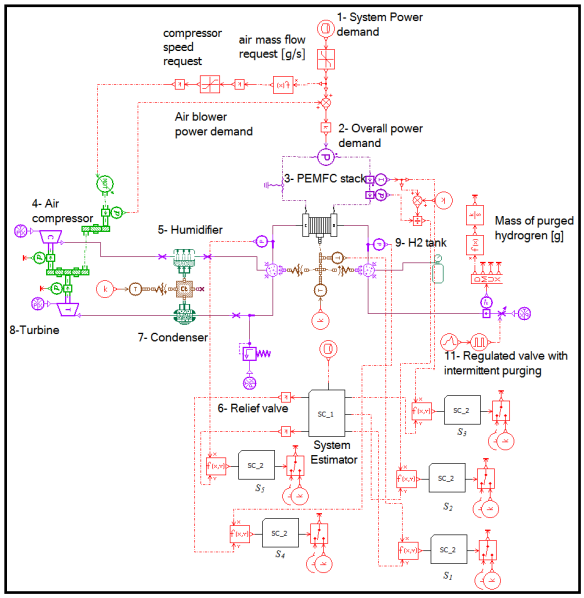


Figure 34: Fault diagnosis system based on diagnostic signals calculated over a window of time.

Note that a super-component was added to this model to help calculate the diagnostic signal over three different windows of time equal to 5 s, 10 s and 15 s. The contents of this super-component for the 5 s moving window of time is presented in Figure 35, whereas the super-component for the 10 s moving window of time is presented in Figure 36, and finally the super-component for the 15 s moving window of time is presented in Figure 37.

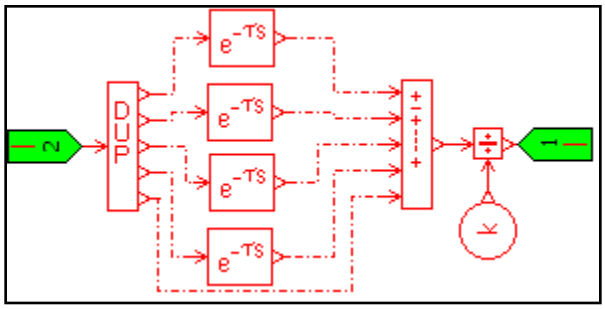


Figure 35: Super-component calculating diagnostic signal over a 5 *s* time window

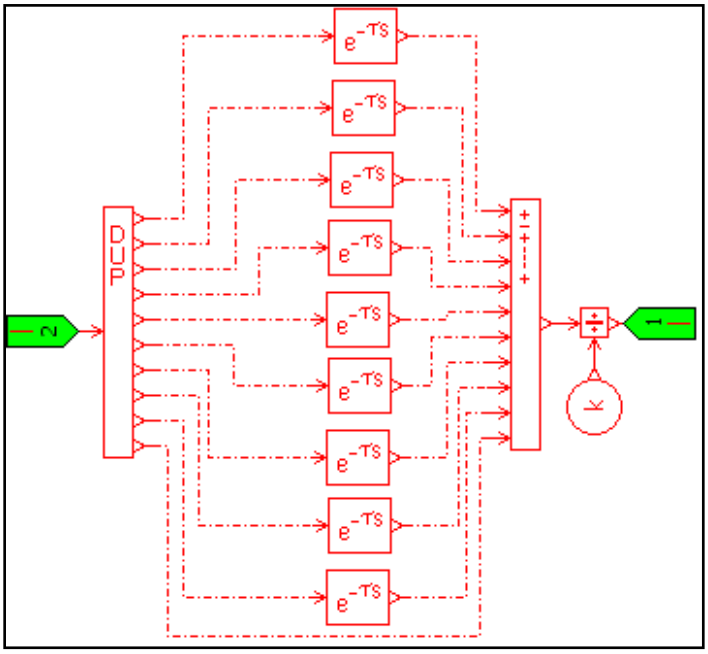


Figure 36: Super-component calculating diagnostic signal over a 10 *s* time window

Furthermore, the effect of the electromagnetic disturbance pulses on the calculation of the five diagnostic signals evaluated using a 5 s moving window of time is presented in Figure 38. Moreover, the effect of the disturbance pulses on the diagnostic signals utilizing a 10 s moving window is presented in Figure 39, and the effect of the disturbance pulses on the diagnostic signals utilizing a 15 s moving window is presented in Figure 40.

Figures 38, 39 and 40 prove that this diagnostic approach of (3) outperforms that of (2). However, it is convenient to mention here that most commercial PEMFC systems commonly have a minimum sampling time of 1 s. Therefore, basing the diagnostic signal on a 5 s window of time can prove to be impractical as seen in Figure 38. Furthermore, the diagnostic signals calculated based on both the 10 s and the 15 s moving windows of time proved to be effective and insensitive to electromagnetic disturbance pulses as seen from Figures 39 and 40. Nevertheless, using a 10 s moving window of time is preferred over the 15 s moving window of time in order to avoid delays in detecting faults as well as saving memory.

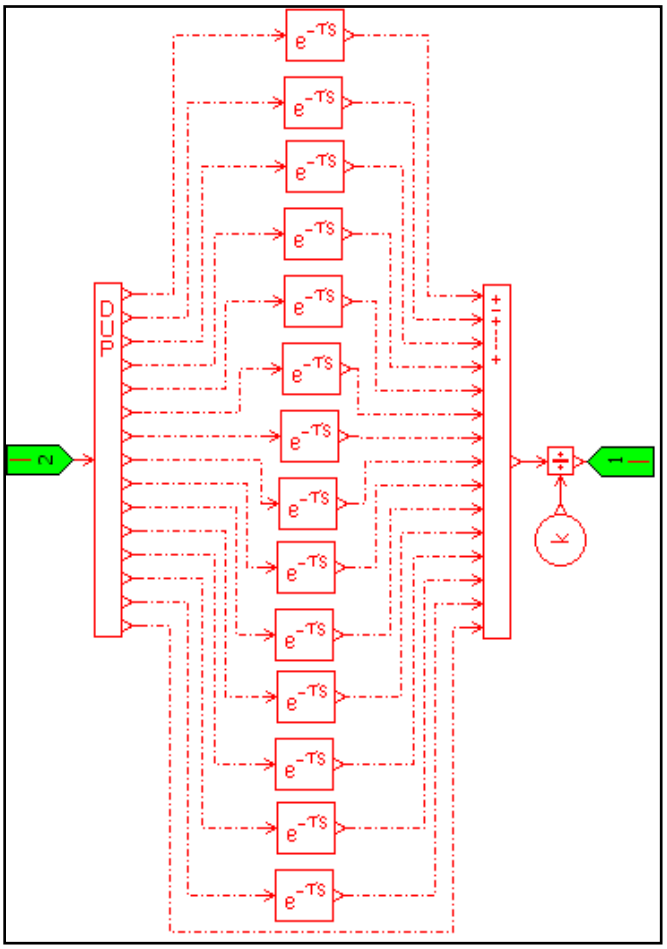


Figure 37: Super-component calculating diagnostic signal over a 15 *s* time window

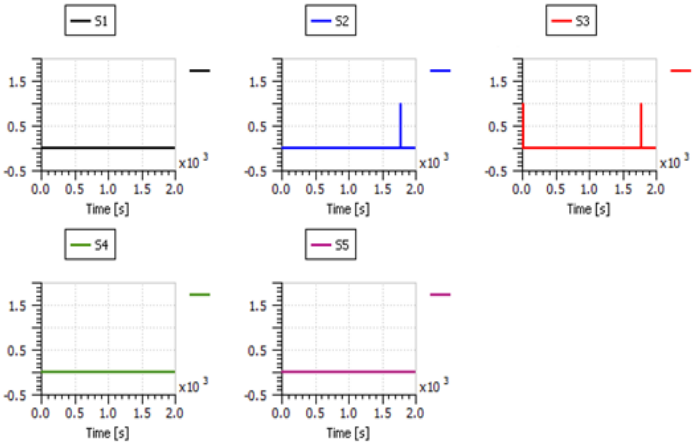


Figure 38: The effect of the disturbance pulses on the five diagnostic signals calculated using a 5 s moving window.

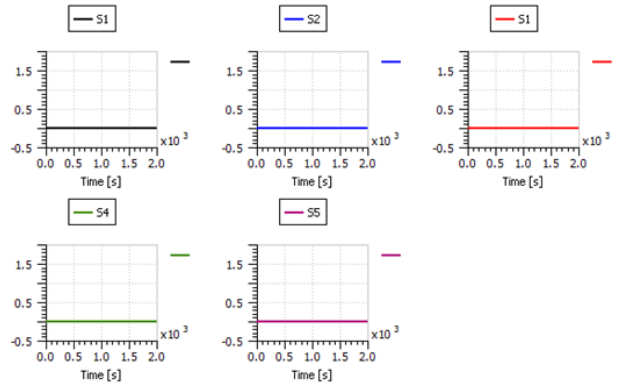


Figure 39: The effect of the disturbance pulses on the five diagnostic signals calculated using a 10 s moving window.

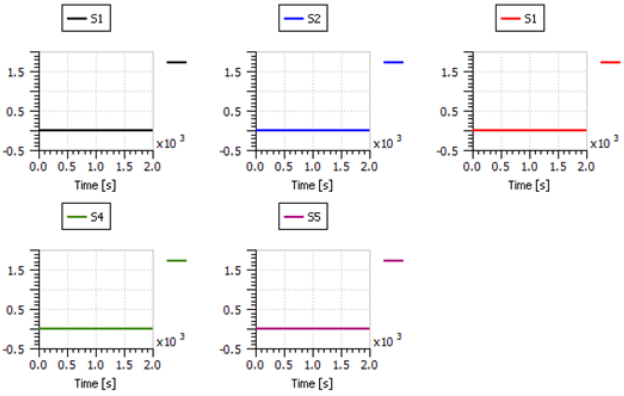


Figure 40: The effect of the disturbance pulses on the five diagnostic signals calculated using a 15 s moving window.

4.3. Fault diagnosis results

To test the proposed fault diagnosis scheme designed using diagnostic signals calculated over a 10 s moving window of time; the five system faults were induced at different times to evaluate the system's ability to detect and isolate the five system faults.

The first fault to be simulated was membrane drying. Thus, the relative humidity target of the humidifier was dropped to 10% and the effect of this action was immediately captured by the diagnostic signals S_2 and S_3 as seen in Figure 41, which – from Table 2 – clearly indicates the presence fault f_1 (drying of the membrane).

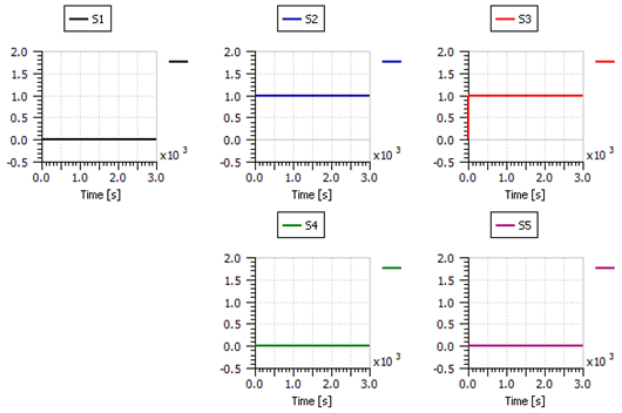


Figure 41: The diagnostic signals successfully detecting drying fault.

The second fault to be tested was membrane flooding. Therefore, the relative humidity target of the humidifier was raised to 100% while reducing the stack temperature to 25°C. Again, this action was clearly reflected on the diagnostic signals as seen from Figure 42, and the three diagnostic Signals S_2, S_3 and S_5 were triggered. However, it was noticed that the effect on the diagnostic signal S_5 was not as fast as that on diagnostic signals S_2 and S_3. This is expected since the increase in pressure drop needs some time to take effect. From Table 2, triggering S_2, S_3 and S_5 at the same time indicates the presence fault f_2 (flooding of the membrane).

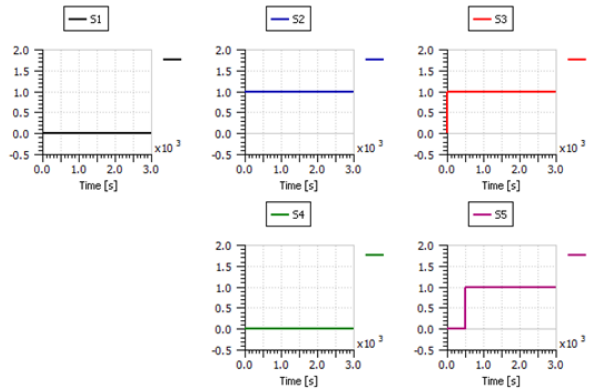


Figure 42: The diagnostic signals successfully detecting flooding fault.

The third fault to be tested was the air leakage in the supply manifold, which was simulated through the addition of a relief valve after the humidifier as shown in Figure 43. This relief valve was set to open with a flow rate of 9 g/s after 500 s of operation. Only one diagnostic signal was affected by this action at around 500 s as depicted in Figure 44. From Table 2, it can be deduced that triggering only S_5 indicates the occurrence of fault f_3 (air leakage).

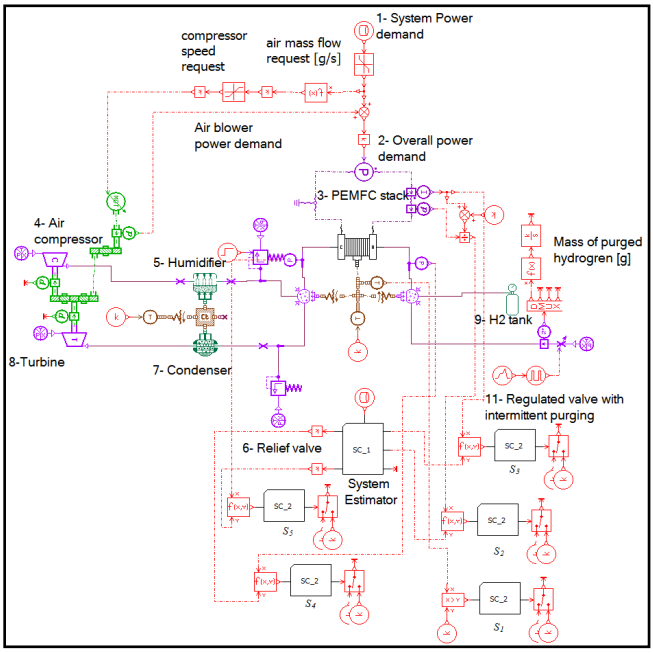


Figure 43: The modified diagnosis model with a relief valve to simulate air leakage.

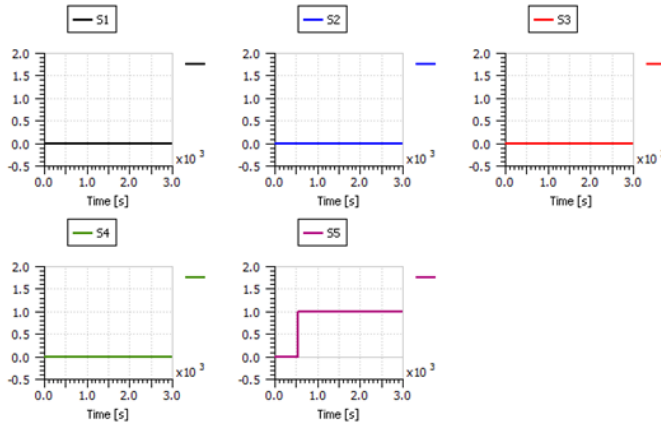


Figure 44: The diagnostic signals successfully detecting air leakage fault.

The next fault to be simulated was hydrogen leakage. As shown in Figure 45, the leakage was simulated through the addition of a relief valve right after the hydrogen supply canister. This valve was set to open with a flow rate of 8 g/s after 500 seconds of operation. The amount of hydrogen leakage can be seen in Figure 46.

From Figure 47, it is noticed that three diagnostic signals were triggered at around 500 s, which are S_2, S_3 and S_4. This clearly indicates the occurrence of fault f_4 (hydrogen leakage) as it can be deduced from Table 2.

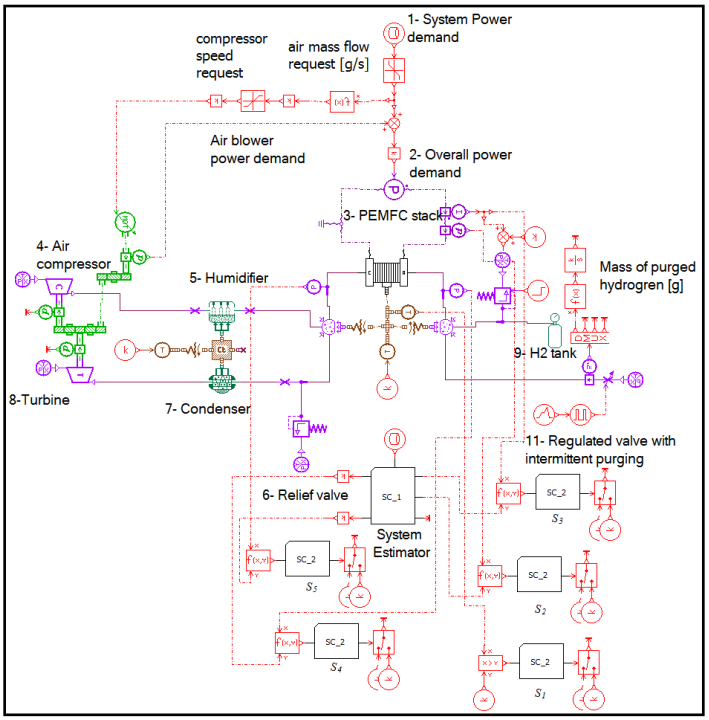


Figure 45: The modified diagnosis model with a relief valve to simulate hydrogen leakage.

Finally, the cooling system failure was tested by forcing the stack temperature to rise to 75°C after 1500 s of operation. The three diagnostic signals S_1, S_2 and S_3 were triggered by this action at around 1500 s of operation as shown in Figure 48. As it can be deduced from Table 2, and as previously explained in section III, S_1 alone is enough to flag fault f_5 (cooling failure) regardless of the effects on other diagnostic signals.

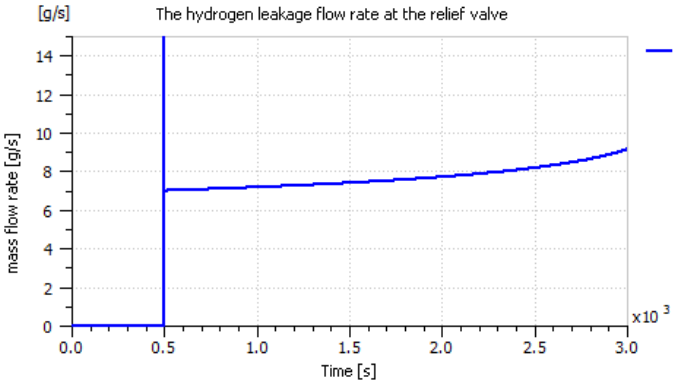


Figure 46: The hydrogen leakage flow rate at the relief valve.

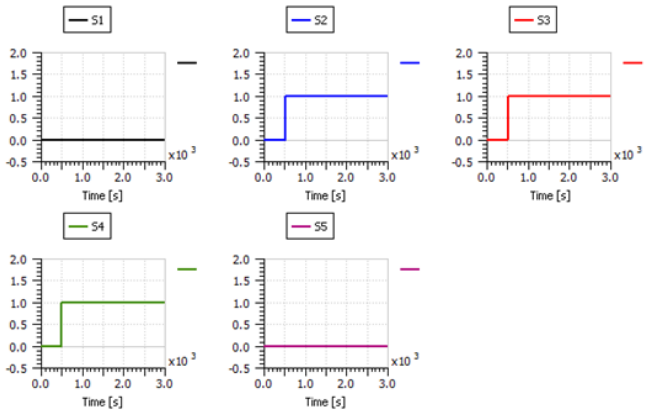


Figure 47: The diagnostic signals successfully detecting hydrogen leakage fault

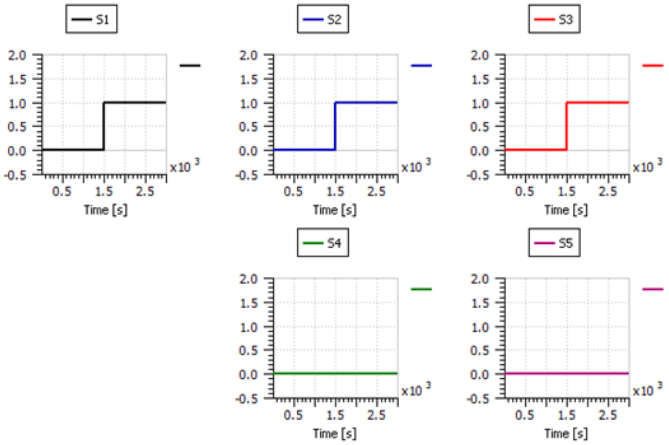


Figure 48: The diagnostic signals successfully detecting cooling system failure fault.

5. Conclusion

Fuel cells are extremely attractive clean power generation systems with the capability of someday replacing fossil fuels in the areas of power generation and transportation, while helping clean the environment by significantly lowering the world's pollution rates. However, to turn this green technology dream into reality, an accurate model that can effectively predict the fuel cell's performance in different conditions is desired. Such model can

then be used to study, simulate, and monitor the behavior of PEMFCs to detect any potential faults that can affect their performance.

Moreover, the complexity of the PEMFC model makes it very difficult and mathematically demanding to try and identify the modeling parameters. Furthermore, other limitations such as the absence of some parameters and confidentiality issues with the manufacturer can also limit the researchers' ability to develop an accurate fault diagnosis oriented model for a commercial PEMFC system. The Siemens software LMS AMESim 14.2 was used in this work as a solution to overcome such limitations.

A diagnosis oriented model of the ElectraGen™ PEMFC system was developed in LMS AMESim and five system faults (drying of the membrane, flooding of the membrane, air leakage, hydrogen leakage in the supply manifold, and cooling failure) were simulated to analyze their effect on different system parameters.

Diagnostic signals based on two different residual generation techniques were also assessed in this work, and the outperforming technique was implemented in the proposed diagnosis scheme. This diagnosis scheme was then tested in LMS AMESim against the five system faults under study and it was found to be very successful in both fault detection and discrimination.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

This paper is part of a joint research program funded by United Arab Emirates University (UAEU) and Japan Cooperation Center, Petroleum (JCCP).

References

[1] R. I. Salim, H. Noura and A. Fardoun, "Fault diagnosis of a commercial PEM fuel cell system using LMS AMESim," in 7th International Conference on Modeling, Simulation and Applied Optimization, Sharjah, UAE, April, 2017. DOI: 10.1109/ICMSAO.2017.7934890

[2] M. Nehrir and C. Wang, Modeling and Control of Fuel Cells: Distributed Generations Applications, IEEE Press Series on Power Engineering, Wiley, 2009.

[3] EG&G Services, Inc., Fuel Cell Handbook, 7th ed., Science Applications International Corporation, DOE, Office of Fossil Energy, National Energy Technology Laboratory, 2004.

[4] T. A. Zawodzinski, M. Neeman, L. O. Sillerud and S. Gottesfeld, "Determination of water diffusion coefficients in perfluorosulfonate ionomeric membranes," J. Phys. Chem., vol. 95, no. 15, p. 6040–6044, 1991. DOI: 10.1021/j100168a060

[5] T. Springer, T. Zawodzinski and S. Gottesfeld, "Polymer electrolyte fuel cell model," J. Electrochem. Soc., vol. 138, no. 8, p. 2334–2342, 1991. DOI: 10.1149/1.2085971

[6] P. Khanungkhid and P. Piumsomboon, "200W PEM fuel cell stack with online model-based monitoring system," Eng. J., vol. 18, no. 4, p. 13 – 26, Oct. 2014. https://doi.org/10.4186/ej.2014.18.4.13

[7] R. I. Salim, H. Noura and A. Fardoun, "A Review on fault diagnosis tools of the proton exchange membrane fuel cell," 2013 Conference on Control and Fault-Tolerant Systems (SysTol), pp. 686 - 693, Nice, France, Oct. 9-11, 2013. DOI: 10.1109/SysTol.2013.6693877

[8] J. Zhang, C. Song and J. Zhang, "Accelerated lifetime Testing for Proton Exchange Membrane Fuel Cells Using Extremely High Temperature and Unusually High Load," J. Fuel Cell Sci. Technol., vol. 8, no. 5, p. 051006, 2011. DOI: 10.1115/1.4003977

[9] C. A. Wilkie, J. R.Thomsen and M. L. Mittleman, "Interaction of poly (methyl methacrylate) and nafions," J. Appl. Polym. Sci., vol. 42, no. 4, p. 901–909, 1991. DOI:10.1002/app.1991.070420404

[10] S. R. Samms, S. Wasmus and R. F. Savinell, "Thermal stability of nafion in simulated fuel cell environments," J. Electrochem. Soc., vol. 143, no. 5, p. 1498–1504, 1996. DOI: 10.1149/1.1836669

[11] N. Ramaswamy, N. Hakim and S. Mukerjee, "Degradation mechanism study of perfluorinated proton exchange membrane under fuel cell operating conditions," Electrochimica Acta, vol. 53, no. 8, p. 3279–3295, 2008. https://doi.org/10.1016/j.electacta.2007.11.010

[12] J. Korbicz, J. Koscielny, Z. Kowalczuk and W. Cholewa, Fault Diagnosis. Models, Artificial Intelligence, Applications, Springer, 2004.

[13] J.-H. Wee, "Applications of proton exchange membrane fuel cell systems," Renewable and Sustainable Energy Reviews, vol. 11, no. 8, p. 1720–1738, 2007. https://doi.org/10.1016/j.rser.2006.01.005

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 310-316 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Parametric Study of Micro Strip Patch Antenna Using Different Feeding Techniques for Wireless and Medical Applications

Debajyoti Chatterjee, Anjan Kumar Kundu*

Department of Radio Physics and Electronics, University of Calcutta, Kolkata-700009, India

ARTICLE INFO

Article history:
Received: 31 October, 2017
Accepted: 09 January, 2018
Online: 30 January, 2018

Keywords:
Micro Strip Patch Antenna
Feeding Mechanism
Return Loss
VSWR
Bandwidth
Dielectric Substrate

ABSTRACT

Due to the increasing bandwidth requirement of modern wireless communication systems, developing antenna having wider bandwidth have been receiving significant attention in the recent years. In this paper, a comparative analysis of the contacting feed (micro strip line and coaxial probe) and non-contacting feed mechanism (both Aperture Coupling and Proximity Coupling) in micro strip patch antenna has been done. In case of contacting feed, RF power is fed directly to the radiating patch using a connecting element such as a micro strip line whereas in case of non-contacting feeds, electromagnetic field coupling is done to transfer power between the micro strip line and the radiating patch. As per the latest research, ultra wide band technology is used in the frequency range from 3 GHz to 10 GHz. We have analyzed and compared the return loss and corresponding bandwidth of these four types of antenna at 5.853GHz so that this antenna may be used in medical as well as wireless applications.

1. Introduction

Micro strip patch antennas are considered as an indispensable tool in today's research oriented activities. The design and manufacturing cost of micro strip antenna is very cheap because of its 2D geometrical structure [1]. The patch antenna is a popular resonant antenna used for microwave wireless communications that require semispherical coverage. Some patch antennas avoid using a dielectric substrate and suspend a metal patch in the air above a ground plane using dielectric spacers; the resulting structure provides increased bandwidth. With an increase in frequency, the input impedance moves to the clockwise direction on the Smith chart [2]. Wireless and Medical applications requires small, low-cost, low profile antennas which has Omni directional radiation pattern in horizontal planes. Micro Strip patch antenna meets all requirements. Figure 1 shows a typical structure of a rectangular micro strip antenna.

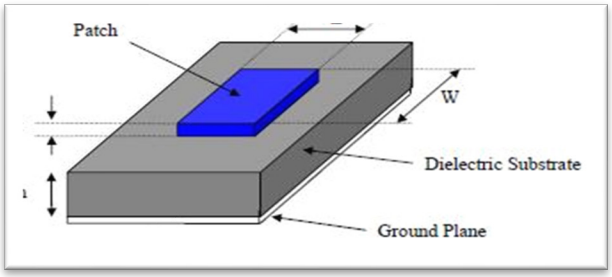


Fig 1: Micro Strip Patch Antenna

2. Feeding Mechanism

Suitable feeding technique plays an important role for antenna efficiency and better impedance matching. The feeding techniques used in the micro strip antenna are given below:

*Dr. Anjan Kumar Kundu, Institute of Radio Physics & Electronics,92, A.P.C Road, Kolkata-700009, +91-9874191584 (M), anjankumarkundu@gmail.com

2.1. Contacting Feed:

In this method, contacting element such as micro strip line or coaxial line is used to help the patch so that it can be fed directly to RF power. The most commonly used contacting feed methods are [3]:

1. Line Feed
2. Co-axial probe Feed

- **Micro Strip Line Feed:**

Micro strip line feed is one of the easier methods to fabricate as it is just a conducting strip connecting to the patch and therefore can be considered as extension of patch.[3-4] It is simple to model and easy to match by controlling the inset position. However the disadvantage of this method is that as substrate thickness increases, surface wave and spurious feed radiation increases which limit the bandwidth.

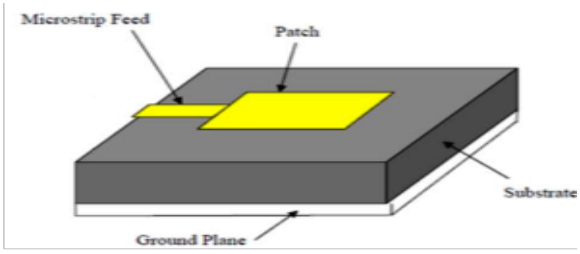


Fig 2: Micro strip Line feeding

- **Coaxial Probe Feed:**

Coaxial feeding is feeding method in which that the inner conductor of the coaxial is attached to the radiation patch of the antenna while the outer conductor is connected to the ground plane.

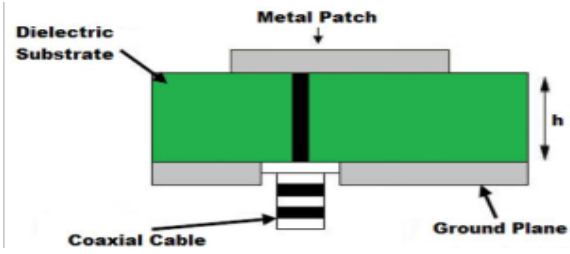


Fig 3: Coaxial Probe feeding

2.2. Non-Contacting Feed:

In this method, the RF power is transferred to the path from the feed line through electromagnetic coupling instead of feeding directly. The commonly used non-contacting feed methods are:

1. Aperture Coupled feed
2. Proximity Coupled feed

- **Aperture Coupling Feed:**

In the aperture coupled feed technique, the radiating patch and the micro strip feed line are separated by the ground plane. Coupling between the patch and the feed line is made through a slot or an aperture in the ground plane.

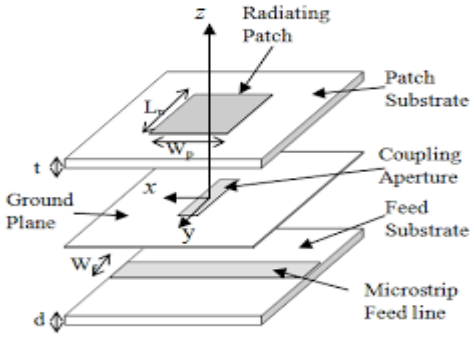


Fig 4: Aperture Coupled feeding

- **Proximity Coupling Feed:**

In this method, two dielectric substrates are placed such that the feed line lies in between the two substrates and the radiating patch is placed at the top of the upper substrate.

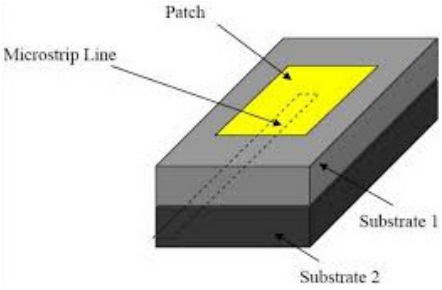


Fig 5: Proximity Coupled feeding

The two dielectric substrates can be selected independently to optimize both micro strip guided waves and patch radiating waves. The ranges of operating thickness of the substrate have a big effect on the resonant frequency and bandwidth of the antenna. Bandwidth of the micro strip antenna will increase with increasing substrate thickness. However, certain limits must not be exceeded, and otherwise the antenna will stop resonating. Therefore, [4], the measures for selecting a substrate may include the following:

(a) Surface-wave excitation.

(b) Dispersion of the dielectric constant and loss tangent of the substrate.

(c) Anisotropy in the substrate.

(d) Cost Effective

3. Study of Antenna Designing Parameters

There are three essential parameters for design of a rectangular micro strip Patch Antenna. Firstly, the resonant frequency (f_0) of the antenna must be selected appropriately. The frequency range for ultra wide band applications is from 3.1GHz to 10.6 GHz and the design antenna must be able to operate within this frequency range.

The second important parameter of good antenna design is dielectric substrate ($\in_r$). A thick dielectric substrate having low dielectric constant is desirable. This provides better efficiency, larger bandwidth and better radiation. FR-4 Epoxy which has a dielectric constant of 4.4 for lower substrate and RT –Duroid for upper substrate having dielectric constant of 2.2 and loss tangent equal to 0.009 can be used for new antenna design [5-6]. The other antenna parameters to be considered for design are length of the patch L, width W, height of dielectric substrate h and Loss Tangent. The patch length is considered around $L_g/2$ (L_g - Length of the ground plane) to initiate the radiation. The antenna is typically fed at the diverging edge on the dimension W because it offers sensible Polarization. The antenna parameters can be calculated by the transmission line method as exemplified below [7]:

3.1 Width of the Patch (mm):

Having specified the height of the patch antenna, the first step in the design procedure is to determine the width of the patch. This can be calculated using the following equation:

$$W = \frac{c}{2f_0\sqrt{\in r + 1}/2} \quad (1)$$

Where,

c = Speed of light in free-space (3×10^8 m/s)

f_0= Resonating frequency

3.2 Calculation of Effective Dielectric Constant (ϵ_{reff}):

This is calculated using the following equation:

$$\epsilon_{reff} = (\in r + 1)/2 + (\in r - 1)/2\left[1 + 12\ \frac{h}{w}\right]^{-0.5} \quad (2)$$

Where, h= height of the patch (mm)

W= width of the patch (mm)

3.3 Calculation of Effective Length (L_{eff}):

This is calculated using the following equation:

$$L_{eff} = \frac{c}{2f_0\sqrt{\in reff}} \quad (3)$$

3.4 Calculation of Length Extension (Δ_L):

The length extension is calculated using the following equation;

$$\Delta L = \frac{0.412h\ (\in reff + 0.3)(w/h + 0.264)}{(\in reff - 0.258)(w/h + 0.8)} \quad (4)$$

Where,

L= Patch length extension (mm)

h = height (mm)

W = width of the patch (mm)

3.5 Calculation of Actual Length:

The actual length of the patch antenna is calculated using the following equation;

$$L = Leff - 2\Delta L \quad (5)$$

4. Design and Simulation of Micro Strip Patch Antenna for Different Feeding Methods

4.1. Micro Strip Line Feed:

The antenna is designed at resonating frequency 5.853 GHz of WLAN. It is designed using transmission line model [8]. This section describes the design of rectangular patch antenna satisfying the given specifications:

Table 1: Design specification of Aperture Coupled Patch

Frequency	5.853 GHz
Antenna Dimension	29mm x 41mm
Dielectric Constant (RT Duroid)	2.2

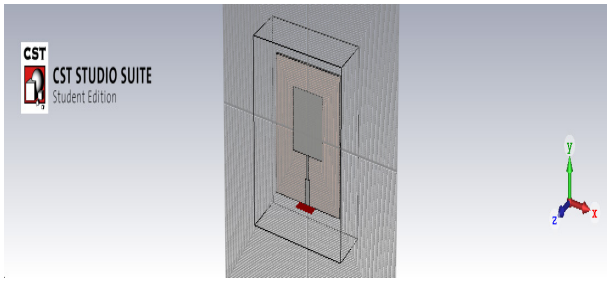


Fig 6: Designed Structures on CST Studio

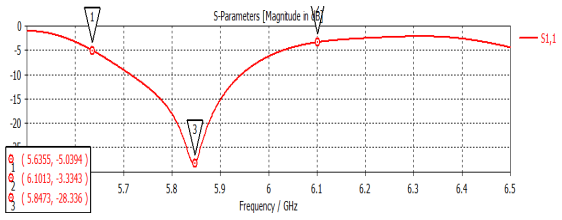


Fig 7: Return Loss S_{11} of simulated antenna at 5.853 GHz

Return Loss:

Fig 7 shows the S11 parameters (Return Loss = - 28dB) for the proposed antenna. The designed antenna resonates at 5.853 GHz.

Smith Chart:

The Smith Chart plot (Fig. 8) represents that how the antenna impedance varies with frequency and gives impedance of 50.4 ohms. For proper matching, the locus must be large so that it connects with the center point of the smith chart.

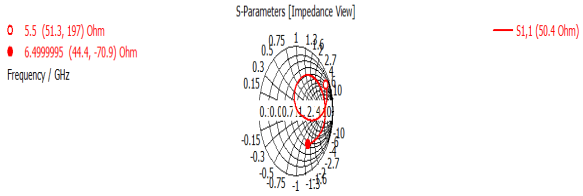


Fig 8: Smith Chart of simulated antenna at 5.853 GHz

Calculation of VSWR:

Fig. 9 shows the VSWR plot against frequency that numerically describes how well the antenna matches with the transmission line it is connected to.

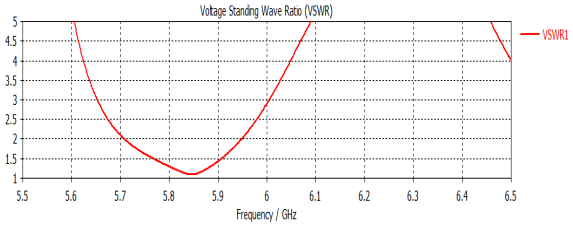


Fig 9: VSWR of simulated antenna at 5.853 GHz

4.2. Coaxial Probe Feeding:

The antenna is designed at resonating frequency 5.853 GHz of WLAN. It is designed using transmission line model. This section describes the design of rectangular patch antenna satisfying the given specifications:

Return Loss:

Fig 11 shows the S_{11} parameters (Return Loss = -37dB) for the proposed antenna. The designed antenna resonates at 5.853 GHz.

Table 2: Design specification of Proximity Coupled Patch

Frequency	5.853 GHz
Antenna Dimension	29mm x 41mm
Dielectric Constant (RT Duroid)	2.2

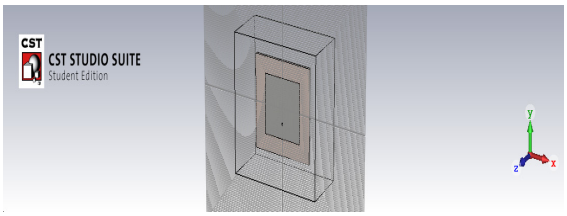


Fig 10: Designed Structure on CST Studio

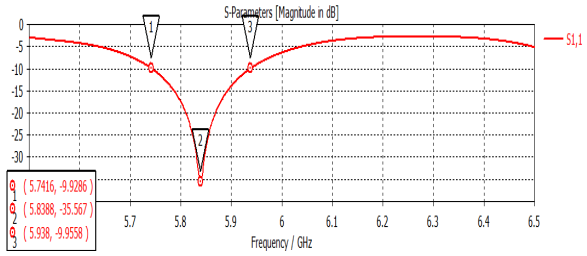


Fig 11: Return Loss S_{11} of simulated antenna at 5.853 GHz

Smith Chart:

The Smith Chart plot (Fig. 12) represents that how the antenna impedance varies with frequency and gives impedance of 50.21 ohms. For proper matching, the locus must be large enough that it passes through the center of the smith chart.

As it can be seen from Fig. 12, the circle cuts the resistive part at 0.5021, thus matching at 50.21 ohm.

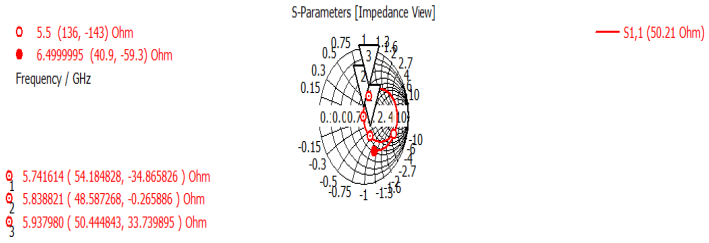


Fig 12: Smith Chart of simulated antenna at 5.853 GHz

Calculation of VSWR:

Fig. 13 shows the VSWR plot against frequency that numerically describes how well the antenna matches with the transmission line it is connected to.

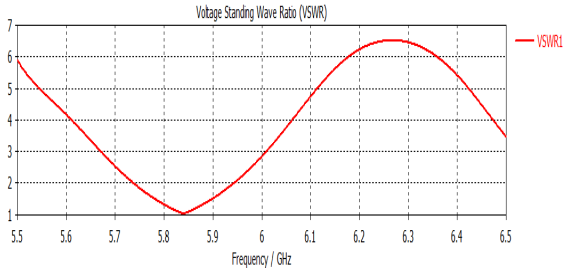


Fig 13: VSWR of simulated antenna at 5.853 GHz

4.3. Aperture Coupled Feeding:

The antenna is designed at resonating frequency 5.853 GHz of WLAN. It is designed using transmission line model. This section describes the design of rectangular patch antenna satisfying the given specifications:

Table 3: Design specification of Aperture Coupled Patch

Frequency	5.853 GHz
Antenna Dimension	29mm x 41mm
Lower Dielectric Constant (FR4)	4.4
Upper Dielectric Constant (RT Duroid)	2.2

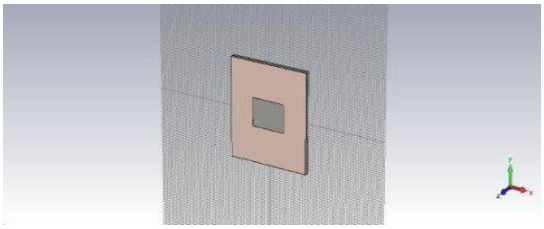


Fig 14: Designed Structures on CST Studio

Return Loss:

Fig 15 shows the S_{11} parameters (Return Loss = - 33dB) for the proposed antenna. The designed antenna resonates at 5.853 GHz.

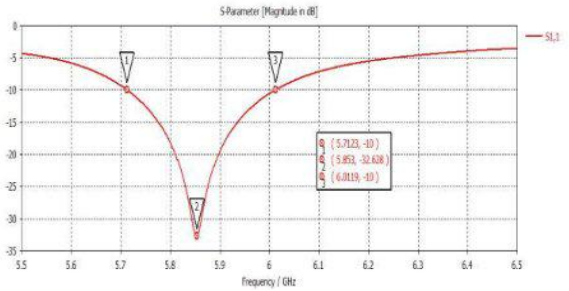


Fig 15: Return Loss S_{11} of simulated antenna at 5.853 GHz

Smith Chart:

The Smith Chart plot (Fig. 16) represents that how the antenna impedance varies with frequency and gives impedance of 49.93 ohms. For proper matching, the locus must be large so that it connects with the center point of the smith chart.

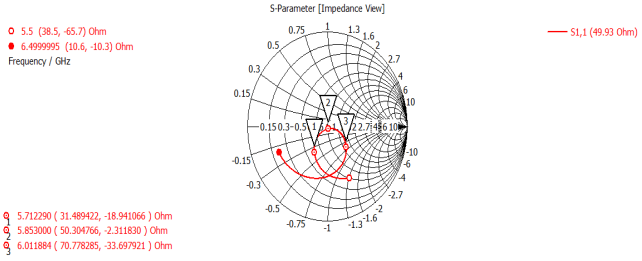


Fig 16: Smith Chart of simulated antenna at 5.853 GHz

Calculation of VSWR:

Fig. 17 shows the VSWR plot against frequency that numerically describes how well the antenna matches with the transmission line it is connected to.

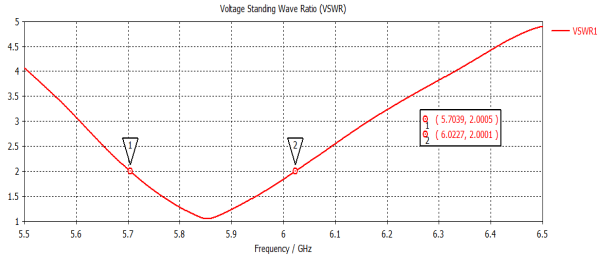


Fig 17: VSWR of simulated antenna at 5.853 GHz

4.4. Proximity Coupled feeding:

The antenna is designed at resonating frequency 5.853 GHz of WLAN. It is designed using transmission line model. This section describes the design of rectangular patch antenna satisfying the given specifications:

Table 4: Design specification of Proximity Coupled Patch

Frequency	5.853 GHz
Antenna Dimension	29mm x 41mm
Lower Dielectric Constant (FR4)	4.4
Upper Dielectric Constant (RT Duroid)	2.2

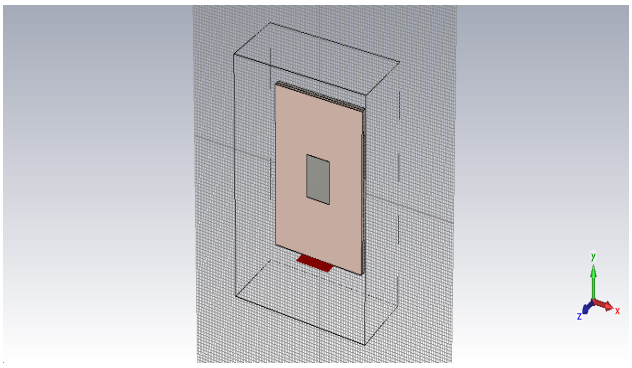


Fig 18: Designed Structure on CST Studio

Return Loss:

Fig 19 shows the S_{11} parameters (Return Loss = -47dB) for the proposed antenna. The designed antenna resonates at 5.853 GHz. More negative the return loss, higher the directivity and gain of the proposed antenna in particular direction.

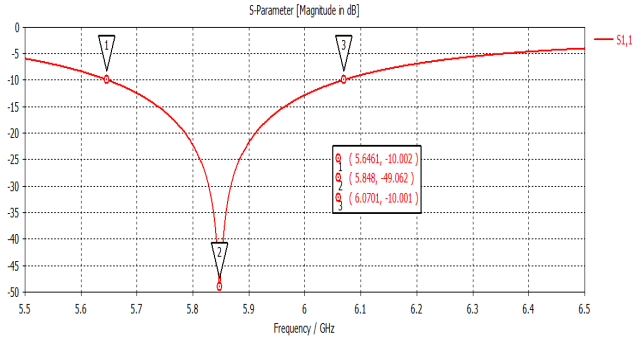


Fig 19: Return Loss S_{11} of simulated antenna at 5.853 GHz

1) Smith Chart:

The Smith Chart plot (Fig. 19) represents that how the antenna impedance varies with frequency and gives impedance of 49.97 ohms. For proper matching, the locus must be large enough that it passes through the center of the smith chart.

As it can be seen from Fig. 19, the circle cuts the resistive part at 0.4997, thus matching at 49.97 ohm.

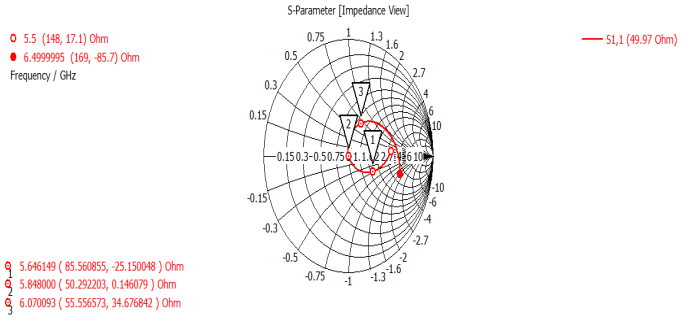


Fig 6.3: Smith Chart of simulated antenna at 5.853 GHz

3) Calculation of VSWR:

Fig. 20 shows the VSWR plot against frequency that numerically describes how well the antenna matches with the transmission line it is connected to.

5. Comparative Study of Important Parameters

Characteristics	Line	Coaxial	Aperture	Proximity
Return Loss (S_{11}) (dB)	-28	-37	-33	-47
Bandwidth (MHz)	250	210	320	475
VSWR	1.08	1.13	1.04	1.16
Impedance (ohm)	50.4	50.21	49.93	49.97

Table E : Comparison of Different Feeding Techniques

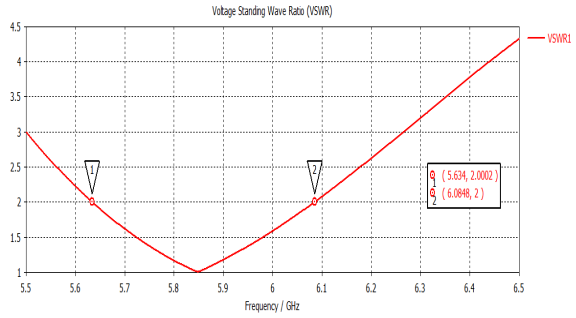


Fig 20: VSWR of simulated antenna at 5.853 GHz

Conclusion

Finally, the optimum result of all four feeding techniques of rectangular patch antenna on FR4 and RT-duroid substrate for Wi-max applications has been investigated. A comparison is made between feeding techniques [9-10] in terms of bandwidth, return loss, VSWR and impedance matching. So, we can see that selection of the feeding technique for a micro strip patch antenna is an important decision because it affects the bandwidth and other parameters also. A micro strip patch antenna excited by different excitation techniques gives different bandwidth, different gain, different efficiency etc. The maximum bandwidth can be achieved by proximity coupling. Proximity coupling gives the best impedance matching and radiation efficiency. Coaxial feeding technique gives the least bandwidth. We can also conclude that by changing the feed point where matching is perfect, the high return loss can be achieved at the resonant frequency. Various micro strip patch antennas with each different feeding technique are presented. The various parameters like return loss, smith chart, and VSWR are plotted for each antenna.

Conflict of Interest

The authors declare no conflict of interest.

Future Scope of Work

The designed patch antenna is intended for wireless applications. This paper mainly deals with the simulation of its characteristics for different feeding techniques which must have to be verified with the fabricated one. The size of this antenna ultimately restricts its usage mainly in biotelemetry in the field of medical applications. So, to extend its usage as an implantable one, its reduced size with similar characteristics has to be investigated. Using bio-compatible ceramic resins instead of conventional one to reduce its size has a great future scope of this work. Also, this design has to be compared with other miniaturized antenna techniques to ensure proper selection of implantable body antenna.

Acknowledgment

The authors of this paper would like to thank University Grant Commission (UGC) for giving the opportunity to work in

Modern Biology Group B: “Image and Imaging” under UPE-II. We also acknowledge all the contribution of corresponding IEEE authors and most importantly the publishers of related books and journals which gave immense support and inspiration in preparing this manuscript. Above all, the extreme mental support and source of inspiration from all the family members and friends are widely acknowledged.

References

[1] C.A.Balanis, Antenna Theory Analysis And Design,Second Edition,John Wiley & Sons. Ramesh Garg,Prakash Bhartie,inder Bahl,Apisak Illipiboon ,Microstrip Antenna Design Handbook, pp.1-68,253-316 Artec House Inc.Norwood,MA

[2] Ramesh Garg,Prakash Bhartie,inder Bahl,Apisak Illipiboon, Microstrip Antenna Design Handbook,pp.168,253-316 Artec House Inc.Norwood.

[3] Amit kumar Jaspreet kaur Rajinder singh,(2013), Performance analysis of different feeding technique,vol 3 issue 3.

[4] K. Praveen Kumar, K. Sanjeeva Rao, T. Sumanth, N. Mohana Rao, R. Anil Kumar, Y.Harish,(2013) Effect of Feeding Techniques on the Radiation Characteristics of Patch Antenna: Design and Analysis International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 2.

[5] Devan Bhalla And Krishan Bansal,(2013) Design of a Rectangular Microstrip Patch Antenna Using Inset Feed Technique IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834,p- ISSN: 22788735. Volume 7, Issue 4 PP 08-1

[6] Brajlata Chauhan, Sandeep vijay, S C Gupta(2013) Comparative analysis of Microstrip Patch Antenna using different substrate and observe effect of changing parameter at 5.4 GHz, Conference on Advances in Communication and Control Systems

[7] Fouzi Harrou, Abdelwahab Tassadit (2010), Analysis and Synthesis of Rectangular Microstrip Antenna ,Journal of Modelling and Simulation of Systems vol 1Issue 1 pp. 34-39.

[8] Rajesh Kumar Vishwakarma, Sanjay Tiwari,(2011) Aperture Coupled Microstrip Antenna for Dual-Band ,Wireless Engineering and Technology vol 2,93-101

[9] John R. Ojha Marc Peters and Igor Mini,(2010) Patch Antennas and Microstrip Lines, microwave and millimeter wavetechnologies modern uwb antennas and equipment ISBN: 978-953-7619-67-1.

[10] Hemant Kumar Varshney, Mukesh Kumar, A.K.Jaiswal, Rohini Saxena and Anil Kumar (2014) Design Characterization of Rectangular Microstrip Patch Antenna for Wi-Fi Application, Vol.4, No.2, E-ISSN 2277 – 4106, P-ISSN 2347 – 5161.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 317-321 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Simulation and FPGA Implementation of a Ring Oscillator Sensor for Complex System Design

Aziz Oukaira[*,1], Idir Mellal[1], Ouafaa Ettahri[1], Mohamed Tabaa[2], Ahmed Lakhssassi[1]

[1]*University of Quebec in Outaouais, Computer Engineering Department, Gatineau, (PQ), 18X 3X7, Canada*

[2]*Moroccan School of Engineering Sciences, LPRI Laboratory, Casablanca, 20250, Morocco*

ARTICLEINFO

Article history:
Received: 10 November 2017
Accepted: 22 January 2018
Online: 30 January 2018

Keywords:
RO
FPGA
VHDL
Complex System Design
DE1

ABSTRACT

This paper, presents the design of a temperature sensor based on RO (Ring Oscillator) in order to make a thermal study for the detection and localization of thermal peaks in a complex system. In this work, a simulation and FPGA implementation of a fully digital temperature sensor features a number of exact inverters that can be dynamically inserted. Before the transition to the implementation in FPGA board, the use of VHDL code is necessary to describe the exact number of inverters that form a single ring oscillator, in order to verify and validate the results obtained. This paper offers a solution to thermally induced stress and local overheating in complex system design which has been a major concern for the designers during the design of integrated circuit. In this paper a DE1 FPGA board cyclone V family 5CSEMA5F31C6 is used for the implementation.

1. Introduction

Decreasing feature sizes and increasing power and package contact densities are making thermal issues extremely important in complex system design. The main purpose of using a temperature sensor is to provide thermal monitoring for predicting local overheating or thermally induced stress. In this article, an intelligent sensor is used for thermal monitoring which is almost an ideal sensor due to its low power consumption. The main idea of this work is to simulate and implement a fully digital temperature sensor that can be dynamically inserted, operated and eliminated from the circuit once the test is done.

The intelligent sensor used in the design is actually based on ring oscillator. One of the important questions in the field of thermal issues of VLSI systems and micro-systems is how to perform the thermal monitoring, in order to indicate the overheating situations, without complicated control circuits.

Traditional approach consists of placement of many sensors everywhere on the chip, and then their output can be read simultaneously and compared with the reference voltage recognized as the overheating level. These techniques, though helpful to reduce the overall power consumption, may cause significant on-chip thermal gradients and local hot spots due to different clock/power gating activities and varying voltage scaling. It has been reported in [1] that temperature variations of 30 °C can occur in a high-performance microprocessor design.

The magnitude of thermal gradients and associated thermo-mechanical stress is expected to increase further as VLSI complex designs move into nanometer processes and multi-GHz frequencies. After the actual temperature is read, the value will be returned through the configuration ports of the FPGA. Then, the sensor will be removed from the chip. This type of oscillator used here is based on the switching time of an inverter. By connecting an odd number of inverters, a naturally oscillating signal is obtained at the output of each inverter of the chain. The oscillation frequency is directly related to the number of inverters. The more inverters there are, the lower the frequency is. To control the oscillation, an inverter can be replaced by a NAND gate, in order to stop or activate the oscillating system. Reference [2], [3], [4], [5] and [6] presents various techniques for varying the frequency of oscillation of the structure. These suggestions based on the voltage control of the delay generated by each cell, on the current control of the rise and falling time of each cell or on the controlled variation of the number of cells.

The advantage of this structure lies in the fact that it can be implemented both with an even number and an odd number of cells. This offers a greater tolerance on the constellation in phase of the signals available at output, but in the case of an implementation with an odd number of cells gives a sinusoidal signal in the output which makes the ring oscillator the perfect solution used ever to

*Corresponding Author: Aziz Oukaira; Email: ouka02@uqo.ca

give more information during the thermal monitoring. In fact, a ring oscillator consists of a feedback loop that includes a necessary odd number of inverters to produce the displacement of the phase which maintains the oscillation the total period is twice the sum of the delays of all the elements that make up the loop. Reversals can be made using the LUT (look-up table) of the configurable logic blocks (CLBs) or the programmable inverters included in the FPGA blocks. In any case, it is useful to insert an external signal to open the loop, as well as an output register to prevent variations in the frequency due to different loads. Thus, the sensor can give instantaneous temperatures.

The future of detectors based on the ring oscillator method is to help designers of more complex integrated circuits to optimize the management of thermal dynamics on the chip [7], [8] and [9]. In this paper in particular, the sensor used allows the detection of thermal peaks. The interests of a sensor based on this method are multiple [10] and [11]. Among the interests of the use of this sensor based on ring oscillator method, is that it can be easily integrated on a chip and can be dynamically inserted or removed from the design at any time due to its small size.

The rest of our work is organized as follows. Section 2, gives a description of the methodology used. Section 3, shows the implementation of the ring oscillator on the FPGA board to verify and validate the results.

2. Description of the methodology used

The new methodology adopted is to validate the control temperature of complex microsystems on a chip based of five inverters forming a ring oscillator [12]. This methodology has given very encouraging results for thermal monitoring in more complex integrated circuits. These simulations and FPGA implementation will then be generalized in high density microsystems, for the development of an integrated thermo mechanical stress control unit using our proposal described in this paper. The proposed ring oscillator depends on the temperature and its frequency changes accordingly. At a given temperature, the oscillator will exhibit a fixed frequency of oscillation.

2.1. Material and geometry of the complex design in COMSOL

We have made the simulation in COMSOL tool appointment with different materials and hardware layers and well on their ranking on the semiconductor. In this figure, we clearly see the reconciliation of our complex circuit layers. The complex model contains 36 Radial Board (RB) in which each RB contains 12 adapter board detector module (ABDM) and each ABDM contains 2 ASICs. Power of 0.6 W is dissipated in a complex circuit of the ASIC (4.68 mm x 5.97 mm) [13]. This amount of power is applied for 9 seconds to visualize the evolution and distribution of heat around the ASIC. Figure 1 shows the ASIC and its support modeled in COMSOL, the ASIC transmits event data through low voltage differential signaling links. To solve the thermal diffusion equations, the Dirichlet boundary conditions (DBC) at 298.15 °K are applied around the daughter board.

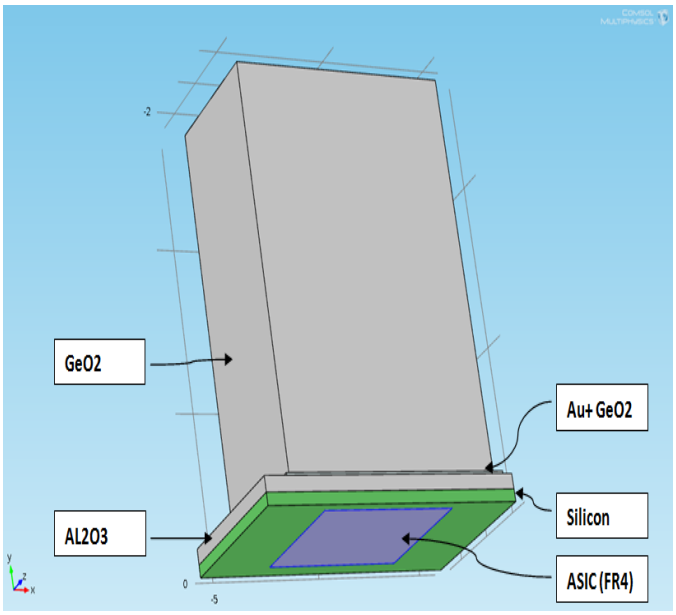


Figure 1 The ASIC modeled by the COMSOL tool.

This structure thus represents a continuous domain, by the method of finite elements consists first of a geometric discretization. The structure is subdivided into sub domains of the simple geometric form called finite element and defined not on the whole of the structure, but for each of its elements.

2.2. Thermal distribution of the complex design in COMSOL

It is very interesting to have a simulation environment that includes the ability to add different physical phenomena to the model studied. In this part, we will present the simulation results from heat sources represented the ASIC complex module in COMSOL tool. As you can see in Figure 2 shows the thermal behavior of our model complex design.

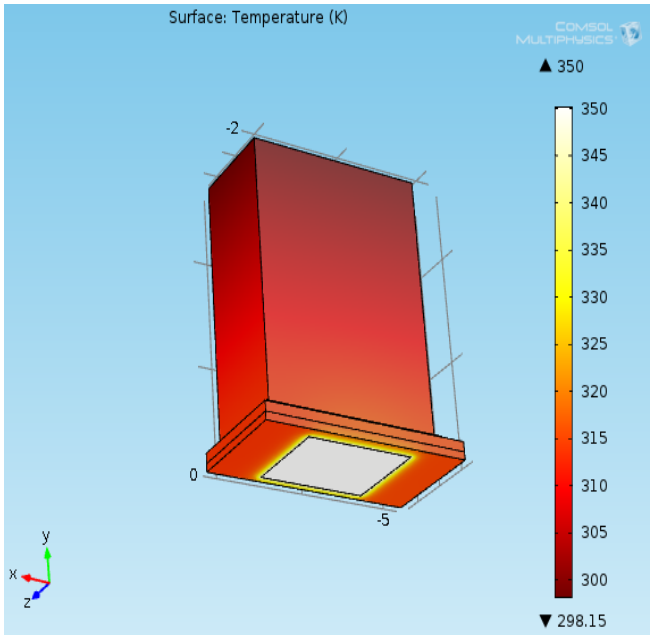


Figure 2 Thermal distribution of the ASIC in COMSOL tool.

This simulation of the ASIC complex module in COMSOL gives a good idea of the behavior and the thermal diffusion of heat sources in our complex system and shows the temperature up to 350 °K. The use of COMSOL tool gives us an idea about the thermal diffusion around the ASIC. According to another study done by [14] and the following Figure 3 presents the values of the temperature as a function of the frequencies of a ring oscillator.

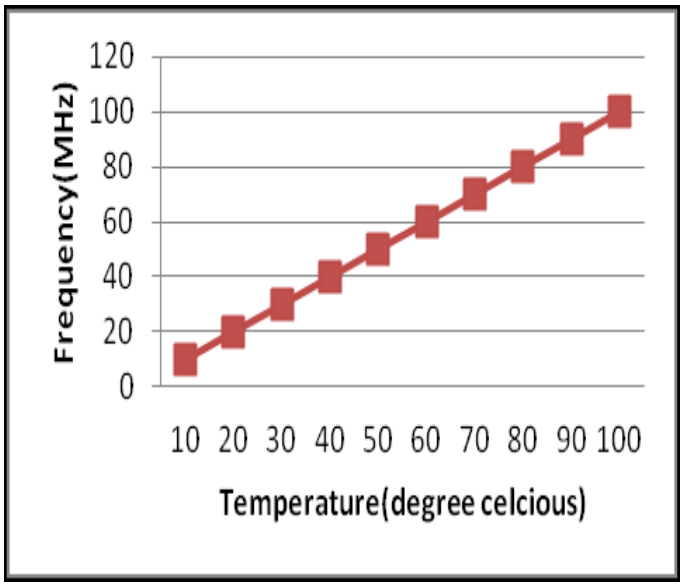


Figure 3 Température of (RO) ring oscillator showing frequency of oscillation linear to temperature.

this research which is done by [14] and the presentation of the results in ghraph form of integrated thermal sensors to monitor the temperature between 10 C to 100 C, according to our simulation under the tool COMSOL Fig. 3 and the study which is done by [12], we can deduce that temperature 77 ° C has our complex module corresponds to approximately 78 MHz (see Fig. 3). This perfectly corresponds to our simulation which is made for the ring oscillator with 5 inverters.

Now we need to validate these theoretical findings find by implementing a 5-inverter based sensor on an FPGA board that allows for different simulations, in [2] explains that more information can be found on the location of the thermal peak at the same frequency and the same temperature of the ring oscillators knowing that this type of ring oscillator sensor can only determine the necessary information if it receives sinusoidal signals, therefore a sinusoidal signal as an input.

3. Experimental implementation and results

The main purpose of this section is the implementation and validation of a single ring oscillator composed of five inverters. VHDL code is used to describe the RO module to facilitate the development of its architecture for its implementation in complex system design. This architecture is modeled in high-level language and simulated to assess its performance and finally implemented on FPGA. The simulation results are validated by using the software Modelsim under Quartus Prime, which allows simulating the behavior of the system in time. Our design flow will be divided into three main parts: simulation, synthesis, and implementation of the VHDL code on FPGA. A description of each part will be presented in the next paragraphs.

3.1. Creation and simulation of the VHDL code

This part, presents the description of the single ring oscillator based of five inverters using a VHDL code editor. The code editor used is Modelsim. Figure 4 shows the top-level module of a single ring oscillator.

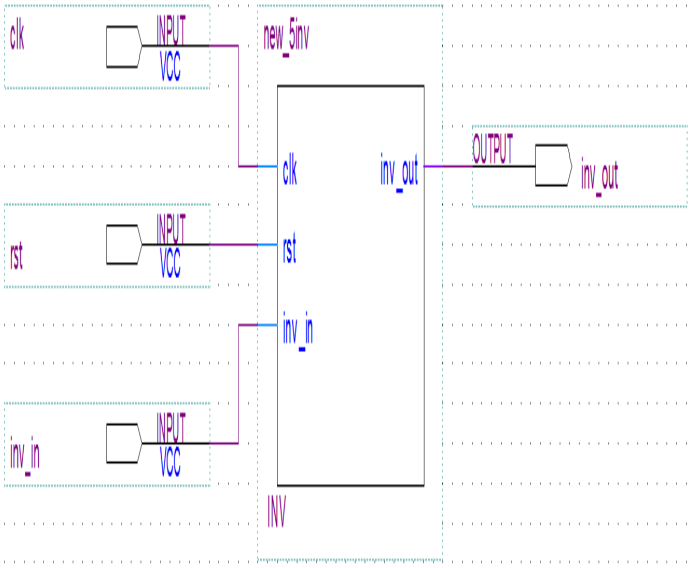


Figure 4 Top level of the ring oscillator composed of five inverters.

This structure in Figure 4 shows the top level of the ring oscillator composed of five inverters, this facilitates the simulation of the logic circuits thereafter. After having followed it, the modelization of a logic scheme of a ring oscillator based on five inverters with Quartus Prime and comes as the next step then the simulation step using ModelSim.

After generating the two .vhd files (the primary file system and the "Test Bench" file) with the "System Generator" the role of the Quartus Prime Navigator comes in order to synthesize the design and generate the RTL files as shown in Figure 5.

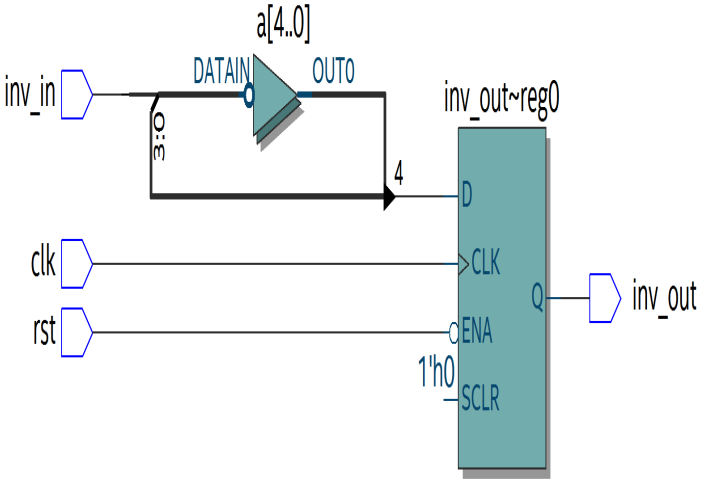


Figure 5 Structure of ring oscillator based five inverters in Quartus Prime tool.

The structure of the single ring oscillator based of five inverters, after synthesis with Quartus Prime from Altera is shown in (Figure

5). The VHDL code implemented was validated. In this part the simulation will be run with the same conditions used in [11], [12] and [14], to validate the experimental results. This Figure 6 shows the results found by the simulation using the Modelsim tool.

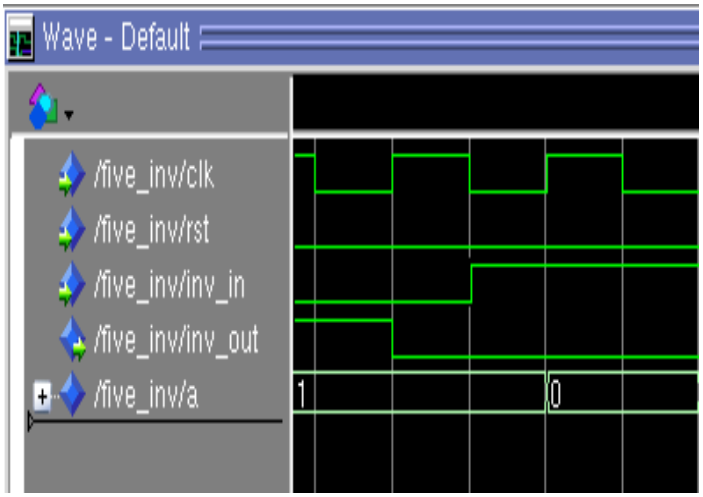


Figure 6 Display résulte of simulation the VHDL code.

As you can see in Figure 6, the signals of simulations are explaining the role of inverters for example for the logical value zero the output is the logical value 1 and the same thing for the second case. The results found validate the VHDL code since it is the correct operation of a single ring oscillator based on five inverters. In this paper, the GDS (gradient Direction Sensor) method for thermal peak detection is used, simulated and verified with a VHDL code and a 'test bench' at the laboratory LIMA the results found meet the initial specifications.

3.2. Implementation and downloading of the VHDL code

Once compiled after the assignment of the pins, the program is ready to be downloaded on the card DE1 cyclone V family and 5CSEMA5F31C6 as a device. This Figure 7 shows that the VHDL code is downloaded successfully on the card.

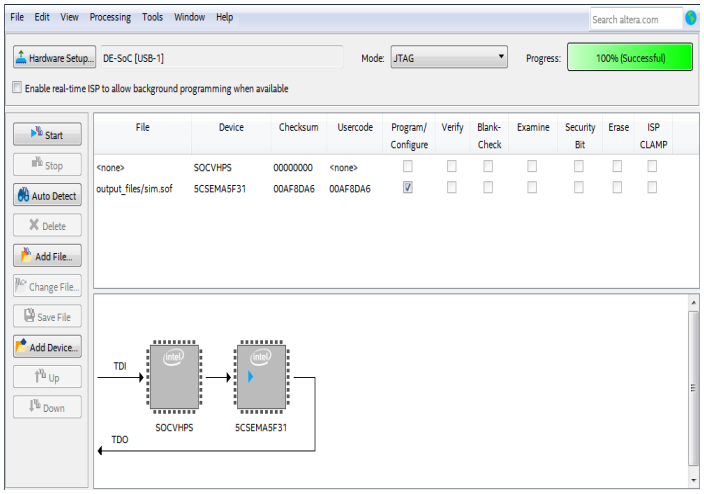


Figure 7 Downloading the code of DE1 Altera cyclone V.

After the download, the program was running and then the outputs were shown on the display of the card. The clock is at 50 MHz, so the outputs should change with a frequency of 50 MHz and the following Figure 8 shows the two values (1 and 0) after implementation on the LCD.

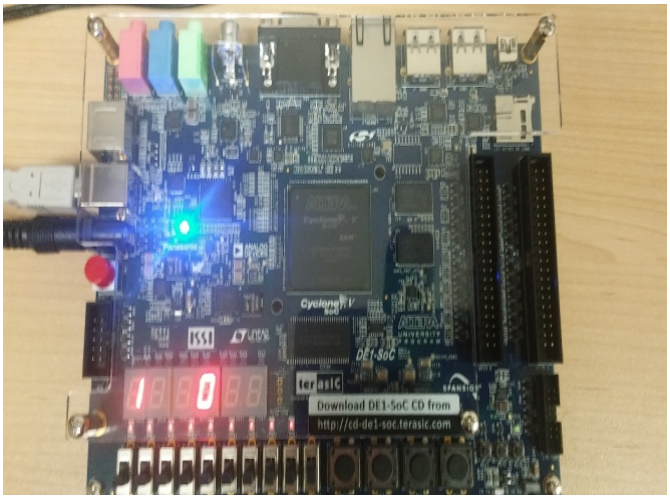


Figure 8 Validation of simulation results on the LCD after the implementation of DE1 Altera cyclone V.

This validation of the simulation results in Figure 8 shows that the value displayed on the LCD matches the results found before. The simulation and implementation on FPGA board DE1 can be applied in any kind of environment to get improved performance to control temperature of complex system design on a chip based of five inverters that will form a single ring oscillator with respect to the conventional schematic; it is also able to keep the temperature constant at the desired value regardless of changes in the load or environment. Thus, the overshooting problem can be solved up to great extent. One of the important issues in the field of electronics is overheating problems especially when it comes to integrated and complex systems and microsystems, but the mean question is how to perform thermal monitoring, to indicate overheating situations, without control.

This type of sensors all over the chip, and then their output can be shown simultaneously and be compared to the reference voltage recognized as the level of overheating. The idea of the proposed method is to validate the results predict the local temperature and gradient along the given distance in some places only on the monitored surface and evaluates obtained several real-time information in a short area in order to predict the temperature of the heat source. These peaks found are essential when monitoring the thermal matrix to avoid a critical induced thermo-mechanical stress. In addition, in most cases, overheating occurs in only one location.

4. Conclusion

The main objective of this paper is to simulate and implement a temperature sensor based on ring oscillator to make a thermal study at the junction. For this paper, we presented an experimental study for the implementation of a fully digital temperature to be dynamically inserted, operated and removed from the circuit after the test. Thus, the main advantage of this type of sensor is the analysis of the temperature during operation of the different blocks of a complex circuit implemented on FPGA. This will be useful for the integrated circuit designer because it offers a solution to thermally induced stress and local overheating in complex system

design which has been a major concern for the designers during the design of integrated circuits.

References

[1] P. Gronowski et al., "High performance microprocessor design," IEEE J. Solid-State Circuits, vol. 33, pp. 676– 686, 1998.

[2] Oukaira, A, Ettahri, O and Lakhssassi, A "Modeling and FPGA implementation of a thermal peak detection unit for complex system design", (IJACSA) International Journal of Advanced Computer Science and Applications, vol.8, no.6, pp.307-312, 2017.

[3] O. Slattery, D. O' Mahoney, E. Sheehan, and F. Waldron, "Sources of Variation in Piezoresistive Stress Sensor Measurements," IEEE transactions on components and packaging technologies, vol.27, no.1, pp.81-86, 2004.

[4] M. Banu, "100 khz – 1 ghz Nmos Variable-frequency Oscillator With Analog and Digital control," ISSCC, Solid-State Circuits Conference. Digest of Technical Papers, vol.31, pp.20-21, 1998.

[5] N. Retdian, , S. Takagi, and N. Fujii, "Voltage controlled ring oscillator with wide tuning range and fast voltage swing," IEEE Asia-Pacific Conference, ASIC Proceedings, pp.201-204, 2002.

[6] Oukaira, A, Pal, N, Ettahri, O, Kengne, E and Lakhssassi, A "Simulation and FPGA Implementation of Thermal Convection Equation for Complex System Design", (IREA) International Journal on Engineering Applications, vol.2, no.6, pp.307-312, 2016.

[7] Pinel, Stephane, et al, "Thermal modeling and management in ultrathin chip stack technology", IEEE Transactions on Components and Packaging Technologies, vol.25, no.2, pp.244–253, 2002.

[8] Oukaira, A, Lakhssassi, A, Fontaine, R, & Lecomte, R, "Thermal Model Development for LabPET II Scanner Adapter Board Detector Module", Proceedings of the COMSOL Conference, pp. 1-5, 2015.

[9] Oukaira, A, Fontaine, R, Lecomte, R, & Lakhssassi, A, "Thermal cooling system development for LabPET II scanners by forced convection flow", In New Circuits and Systems Conference (NEWCAS), 15th IEEE International, pp. 289-292, 2017.

[10] Rahmanikia, Navid, et al, "Performance evaluation metrics for ring-oscillator-based temperature sensors on FPGAs: A quality factor", Integration, the VLSI Journal, vol.57, pp.81–100, 2017.

[11] Rahmanikia, Navid, et al, "Exploring Efficiency of Ring Oscillator-Based Temperature Sensor Networks on FPGAs", Proceedings of ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp.264–264, 2015.

[12] Oukaira, A, Mellal, I, Ettahri, O, Kengne, E and Lakhssassi, A "Thermal Management and Monitoring Based on Embedded Ring Oscillator Network Sensors for Complex System Design", (IJCEIT) International Journal of Computer Engineering and Information Technology, vol.9, no.7, 2017.

[13] Oukaira, A, Taheri, S, Nour, M, & Lakhssassi, A, "Simulation and Validation of Thermal Stability for Complex System Design High Power Dissipation", In the 5th IEEE International Conference on Smart Energy Grid Engineering (SEGE), pp. 229-233, 2017.

[14] Suman, S, & Singh, B. P. "Ring oscillator-based CMOS temperature sensor design", International Journal of Scientific & Technology Research, vol.1, no.4, pp.76–81, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 322-328 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Dynamic Reallocation Based Window Access Scheme for Enhancing QoS of Vehicular Ad-hoc Networks (VANETs)

Md. Amirul Islam*, Hossen Asiful Mustafa

Institute of Information and Communication Technology,

Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh

ARTICLE INFO

Article history:
Received: 14 November, 2017
Accepted: 18 January, 2018
Online: 30 January, 2018

Keywords:
VANET
V2I communications
MAC
TDMA

ABSTRACT

This article proposes a new MAC scheme for Vehicle-to-Infrastructure (V2I) communications that dynamically reallocates unused TDMA slots. By maintaining a balanced waiting time, the proposed TDMA based scheduling scheme allocates TDMA slots in a rational way to minimize merging and one-hop neighboring collision. The proposed scheme ensures dynamic reallocation of unused slots by using "time slot reassignment" mechanism. The scheme has been simulated in VEINS framework of OMNET++ network simulator and its performance has been compared with other conventional protocols. Experimental results show that our scheme performs better than existing schemes in terms of successfully transmitted data packets.

1. Introduction

In recent years, Vehicular Ad-hoc Networks (VANETs) technology has drawn interest of many researchers as a common platform for inter vehicle communication on highways or in urban environments [1-5]. As a special type of Mobile Ad-hoc Network (MANET), VANET provides communications among vehicles and between vehicle, and infrastructure via Road Side Units (RSUs). The relevance of VANET has been confirmed by the development of a specific IEEE standard, 802.11p, to support VANETs [2].

Different from other ad-hoc networks, VANET has unique characteristics of high node mobility, dynamic topology changes and strict delay constraints. The applications devised for VANETs can be divided into the following three services: safety services, traffic management and user-oriented services. Among these services, safety services usually require bounded transmission delays as well as low access delays. As the IEEE 802.11p standard does not provide a reliable broadcast mechanism with bounded communication delay [2, 6], it is not sufficient for VANET applications which are primarily envisioned to improve road safety.

As opposed to IEEE 802.11p, Time Division Multiple Access (TDMA) is inherently a collision free scheme with bounded access delay. Several protocols have been proposed in VANETs using TDMA to provide fairness and to reduce interference among vehicles. By concatenating or rescheduling time slots based on access priority, TDMA based schemes can also assign bandwidth resources to different vehicles on-demand. However, for TDMA based Medium Access Control (MAC) protocol, an efficient slot management is very crucial to ensure fairness as well as reusability of unused slots. VANETs, which have very dynamic topology, are very much prone to allocating a time slot in overlapping areas. In this case, two types of collision may occur: access collision between vehicles trying to access the same available time slots, and merging collisions between vehicles using the same time slots. Some researchers worked on collision issues and some worked on fairness; however, a comprehensive protocol to ensure fairness and reallocation of unused slots is still missing.

In this work, we propose a new TDMA based MAC scheme, Dynamic Reallocation based Window Access MAC (DRWA-MAC), to achieve fair slot allocation as well as reallocation of unused slots dynamically. Our scheme allocates time slots to the registered vehicles evenly; then it monitors the usage of the assigned slots to ensure the dynamic reallocation of unused slots in a fair way. Simulation results in OMNET++ show that our proposed scheme outperforms several existing schemes. This work

*Corresponding Author: Md. Amirul Islam & Email: amirulctg@yahoo.com

is an extension of our previous work originally presented and published in ECCE conference [1].

2. Related Work

VeMAC [7, 8] is a TDMA based scheme proposed for inter-vehicle communication in which vehicles in opposite directions (Left, Right) and roadside units (RSUs) are assigned with time slots in the same TDMA time frame by using logical segmentation. Though VeMAC can make use of the seven DSRC channels and decrease the rates of merging, and access collisions, it is burdened with the overhead of the size of the transmission control frame and cannot ensure reusability of an unused slot.

In ACFM [9], each RSU maintains a dynamic slot assignment cycle for vehicles in its coverage adaptively. Under the scenario of light traffic, ACFM controls the excessive increase of unassigned slots by shrinking slots assignment cycle frame by frame. When there is a mass of vehicles on roads, ACFM provides more available slots by expanding cycle. However, to avoid interference between adjacent segments, ACFM requires two orthogonal frequencies to ensure that the same frequency is not used for a distance of two hops.

TC-MAC [6] is a TDMA Cluster-based MAC protocol for VANET. It subdivides slots of control channel (CCH) into mini-slots to broadcast beacons or safety messages. Here, in each frame, each vehicle gets allocation on a time slot in service channel (SCH) and competes for a mini-slot on the CCH. A vehicle uses its mini-slot to inform the other vehicles of its transmission during time slot on the SCH. Cluster members can use their time slots on the service channels to exchange non-safety data in unicast or multicast communication mode. TC-MAC protocol was designed for simple highway traffic in which all the vehicles are moving in the same direction and it has high collision in bidirectional traffic and in urban scenarios due to the merging collision problem. TC-MAC also suffers from inter-cluster interference problem when two or more clusters are in close proximity.

PTMAC [10] is a prediction-based TDMA MAC protocol that predicts encounter collisions and effectively reduces the number of collisions while maintaining high slot utilization. Here, immediate vehicles within a two-hop neighborhood detects a potential collision if two vehicles occupies the same slot and the protocol removes the collision by asking one of them to change its current slot. As PTMAC makes TDMA slot assignment on a contention based approach, it is unable to ensure fairness.

FAWAC-MAC [1] is a TDMA based scheme that uses capture effect to send extra data in free or unused slots. Here, all vehicles, which are not the owner of the current slot, check whether its own slot is within a predefined window size. If not, they transmit their packets in low power while the original owner of the slot transmits packets in high power. Using capture effect, the RSU can distinguish the power levels. In the presence of a high power packet, all low power packets get discarded; while in the absence, low power transmission succeeds. When vehicle density is very high, FAWAC-MAC has a chance of low power packet collision in free slots and if the coverage area is large, then it may also be suffered by near-far effect.

FAWAT-MAC [1] uses time slot reassignment mechanism to transmit data in unused slots. Here, if an RSU detects no ongoing transmission within one fourth of the beginning time of a slot, it selects the slot as unused and invites another vehicle to transmit data in that slot. Upon reception of the invitation packet, the specific vehicle transmits data for the remaining time of the slot in a collision free manner. The slot allocation process of FAWAT-MAC can cause merging and one-hop neighboring collision for boundary vehicles. In this protocol, the waiting time for some of the high speed boundary vehicles can be so high that by the time they got their assigned slot, they may have left the corresponding RSU.

3. Proposed System

In this work, we propose Dynamic Reallocation based Window Access MAC (DRWA-MAC), an extension of the FAWAT-MAC protocol [1]. Our proposed protocol operates as a TDMA based MAC protocol in a centralized topology where RSUs control distribution and reallocation of slots.

3.1. Assumptions

We made the following assumptions regarding VANET's context in which DRWA-MAC operates:

- Each vehicle and RSU has a unique ID.
- Vehicles and RSUs are equipped with GPS and are perfectly time synchronized.
- At the beginning of each frame, the RSU transmits registration beacon and then, each vehicle, which receives that message, broadcasts a registration response message with its own ID, location and data size. For a certain amount of time, the whole registration response system is a contention based process.
- Each newly joined vehicle that does not have a slot and wants to get a new slot shall listen to the channel for one frame. Then, after receiving the registration beacon, they can register to the RSU.

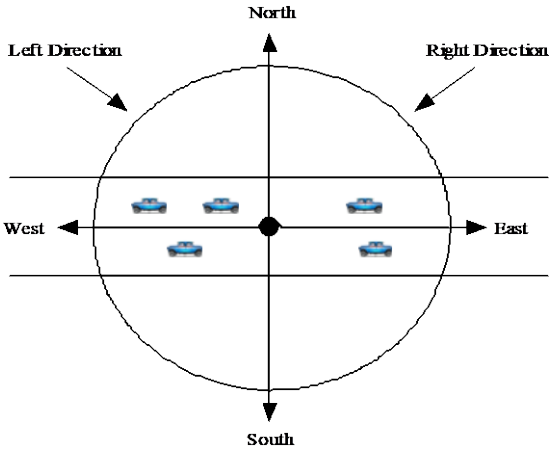


Figure 1: Right and left direction of vehicle movement as defined in the protocol

3.2. System Design

In our proposed DRWA-MAC protocol, we consider RSU based centralized VANET with a set of vehicles moving in lanes consisting of two-way traffic. In a two-way traffic system, we

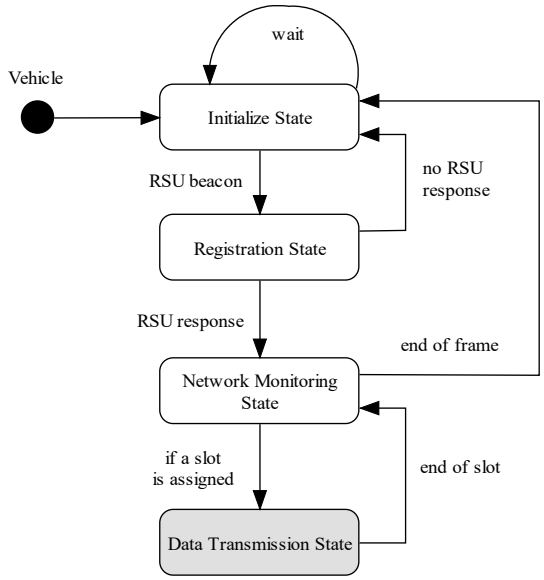


Figure 2: Vehicle state diagram in our scheme where a vehicle can be in one of the four states: initialize, registration, network monitoring, and data transmission

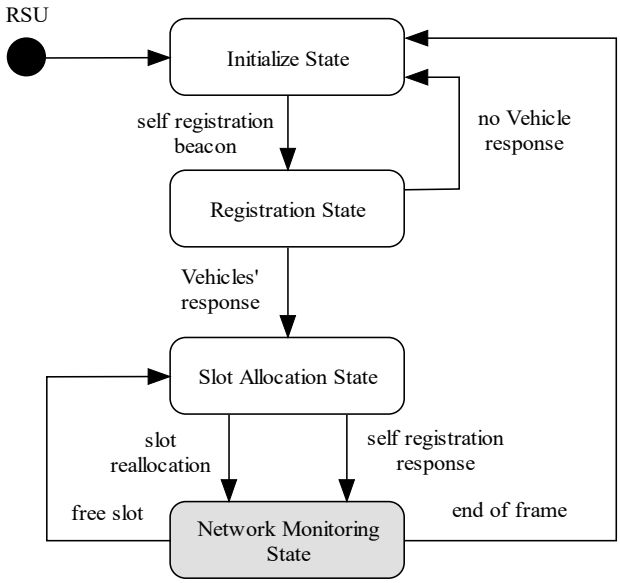


Figure 3: RSU state diagram in our scheme where the RSU can be in one of the four states: initialize, registration, slot allocation, and network monitoring

define a vehicle as positioned in left direction of a RSU's broadcasting area if it is placed in the west half and right direction, if it is placed in the east half, as shown in Figure 1.

Our scheme consists of three processes: (i) a contention based registration process, (ii) a fair slot distribution process and finally, (iii) a network monitoring process for reallocating unused slots. Considering the effect of these processes, the states of a vehicle and RSU in our proposed scheme is depicted in Figure 2 and Figure 3 respectively.

As shown in Figure 2, upon reception of RSU's registration beacon, a vehicle moves into the *registration state* from its *initialize state* and sends back a registration response. Then, after receiving RSU's response, the vehicle moves into the *network monitoring state*; once a slot is assigned, the vehicle moves to the *data transmission state* where it transmits data in the slot assigned to it. At the end of the slot, the vehicle moves back to the *network monitoring state* and waits for additional slots from the RSU. When data transmission is complete, the vehicle moves back to the *initialize state*.

As shown in Figure 3, after transmitting registration beacon, an RSU enters into the *registration state* where it receives response from vehicles for a short amount of time. After receiving registration response(s) from the vehicles, the RSU enters into the *slot allocation state* where it allocates available slot to the vehicles; then, it transmits a registration response message and moves to the *network monitoring state* where the RSU monitors network for free slots. If any free slot is detected, the RSU moves back to the *slot allocation* state; otherwise, it moves to the *initialize* state at the end of the TDMA frame.

4. Details of the DRWA-MAC Protocol

DRWA-MAC consists of three processes: (i) a contention based registration process, (ii) a fair slot distribution process, and (iii) a network monitoring process. In this section, we discuss the details of these processes and also, present our slot distribution algorithm.

4.1. Registration Process

At the beginning of the TDMA frame, RSU transmits a registration beacon for all the vehicles in its broadcasting area. Upon reception of the registration beacon, each vehicle sends back a registration response message to the RSU using a contention based approach. To accommodate the vehicles response, this contention based procedure continues for a certain predefined amount of time. Within registration response message, vehicles send back their ID, location and data size.

4.2. Slot Distribution Process

After registration process, RSU's usually have N number of participating vehicles that need to be assigned with S number of slots. As the registration response from each vehicle consists of location information, RSU can differentiate the L number of vehicles on its left and the remaining (N-L) number of vehicles on its right direction in a road with two-way traffic. Our protocol divides the available S slots into $S/2$ ODD, and S/2 EVEN sets. Then, the $S/2$ ODD slots are distributed among the L left directional vehicles and S/2 EVEN slots to (N-L) right directional vehicles. Our protocol uses a repeated round robin approach for channel assignment. In case the number of vehicles is less than the number of slots, i.e., S>N, one or more vehicle may get multiple slots in the same frame. An example is shown in Figure 4 where there are 8 slots in a frame with two left direction vehicles (*V[1]*, *V[3]*), and two right direction vehicle (*V[2]*, *V[4]*); a repeated round robin distribution will allocate slot 1, and 5 to *V[1]*, slot 2, and 6 to *V[2]*, slot 3, and 7 to *V[3]* and slot 4, and 8 to *V[4]*. On the other hand, if $L < S/2$ and $(N–L) > S/2$, one or more left directional vehicles will get multiple slots but $(N–L–S/2)$ right

Slot Numbers

1	2	3	4	5	6	7	8
V[1]	*V[2]*	*V[3]*	*V[4]*	*V[1]*	*V[2]*	*V[3]*	*V[4]*

Figure 4: Example of slot allocation in our proposed scheme using repeated round robin distribution where the index represents vehicle ID. In this example, there are 4 vehicles where each vehicle received 2 slots each

directional vehicles will not get any slot allocation. Our slot distribution algorithm is shown in Algorithm 1.

Algorithm 1: TDMA Slot Distribution

```
1:  foreach available slot s in frame f
2:      calculate even slot es and odd slot os
3:  foreach registered vehicle v
4:      If vehicle v is right directional and es>0
5:          assign an  even slot es to vehicle v
6:          es = es - 1;
7:      else if os>0
8:          assign an odd slot os to vehicle v
9:          os = os – 1;
10: s = s – v;
```

4.3. Transmission Monitoring Process

After slot assignment, vehicles usually begin transmission in their assigned slots. Even though all the vehicles register for TDMA slot, some of them may not use all of their assigned slot(s); this could be due to early completion of data transmission, passing over RSU's broadcasting area, detection of the safety message that is already broadcasted, etc. To avoid wastage, for each slot, our protocol employs a monitoring process at the RSU end to check whether an existing transmission is going on or not; if not, then it ensures reallocation. Here, an RSU detects transmission status by monitoring the first one fourth of the beginning time of a slot; if no data is transmitted during this period, the RSU mask the slot as a free and broadcasts a very light NULL packet similar to a becon signal. Inside the NULL packet, RSU sends the ID of the vehicle having maximum data size; this is possible because at the beginning of the frame, each vehicle registers with its data size. Upon reception of the NULL packet, the specific vehicle transmits data for the remaining time of the slot and hence, reallocation of unused slots is achieved in a collision free manner.

Figure 5 shows operation of the proposed DRWA-MAC in detail. Here, *V[1]*, and *V[3]* are left directional vehicles and *V[2]*, *V[4]* are right directional vehicles. *V[1]*, *V[2], V[3], and V[4]* have 24MB, 22MB, 4MB, and 4MB of data to transmit respectively. These vehicles V[1], V[2], V[3], V[4] are assigned with slots (*1*,*5*), (*3*,*7*) and (*2*,*6*), (*4*,*8*) respectively by our slot allocation algorithm. Considering max data size of each slot as *4* MB, at slot *7*, RSU will detect that there is no ongoing transmission in channel as *V[3]* has no data left to send and therefore, will select

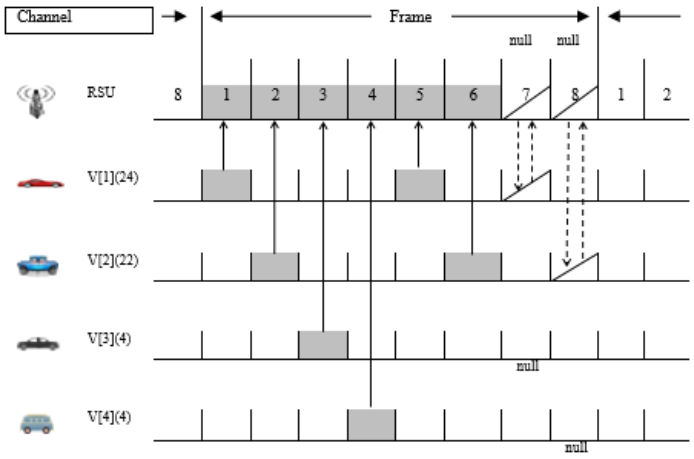


Figure 5: Operation of proposed DRWA-MAC with four vehicles and 8 slots in each frame where 4MB of data can be transmitted in each slot. Initial slot assignment is as shown in Figure 4

V[1] as it has *16* MB data remaining for transmission (out of 24 MB, V[1] has transmitted 8 MB in slot 1 and 5). So, *V[1]* will be able to send 3 MB of data in the remaining 3/4 of the slot and will have 13 MB data left for transmission. At slot *8,* RSU again will detect that there is no ongoing transmission in the channel as *V[4]* has no data left to transmit and will select *V[2]* as it has 14 MB data remaining while *V[1]* has 13 MB of data remaining to transmit.

5. Effectiveness

In our protocol, the use of "time slot reassignment" technique makes use of most of the wasted slots and hence, increases overall throughput of the TDMA network. As our protocol works in RSU based centralized topology, it is free of access collision. Since the proposed protocol uses RSU's guidance based "time slot reassignment" technique where a single vehicle is selected by RSU to send message in free slots, it is also free from hidden node collision.

Instead of straight forward segmentation like VeMAC, our protocol divides and allocates slots in ODD-EVEN basis and it balances waiting time of the partitioned vehicles. This distribution process also avoids or minimizes merging collision and one-hop neighboring collision. To support our argument, an example scenario is depicted in Figure 6. In this scenario, there are 2 RSUs, RSU[0], and RSU[1]. We show that vehicle *V[4]*, *V[6]* of *RSU[0]* and *V[5]*, *V[7]* of *RSU[1]* are one hop neighboring collision free as *RSU[0]* assigns EVEN slots to *V[4]*, and *V[6]* and on the other hand, *RSU[1]* assigns ODD slots to the vehicles *v[5],* and *v[7]*. Additionally, if we consider the position of vehicle *V[5]* of *RSU[1]* and *V[1]* of *RSU[0],* we can claim that our protocol also minimizes merging collision because both *V[5]* and *V[1]* are placed in left/west half of their respective RSU and each is assigned with ODD slots. By the time *V[5]* enters into the coverage area of *RSU[0]*, *V[1]* may leave and the slot freed by *V[1]* can be reassigned to *V[5]* so that *V[5]* can transmit data in the running frame.

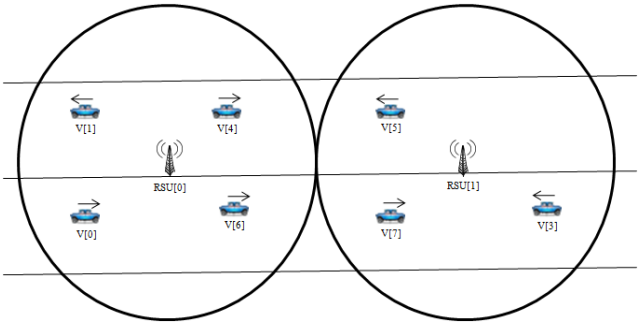


Figure 6: Example scenario of proposed DRWA-MAC protocol which depicts that the protocol is free from one-hop neighboring collision and merging collision

6. Simulation Results and Analysis

In our simulation experiments, we have compared the performance of our proposed DRWA-MAC protocol with ACFM, VeMAC, PTMAC, FAWAC-MAC and FAWAT-MAC in terms of number of successfully transmitted data packet, average delay, and data loss ratio on overall road traffic network.

6.1. Simulation Environment

We have implemented DRWA-MAC protocol on the VEINS [11] framework of OMNET++ [12] network simulator. The vehicle movement and related realistic road traffic scenario is generated using SUMO [13] urban mobility traffic generator.

In our scenarios, we considered a 2000m×1000m rectangular two way road traffic network with ten junctions as shown in Figure7, where each junction contains three lanes. We have covered our road network segment with three non-overlapping RSUs as shown in Figure 8.

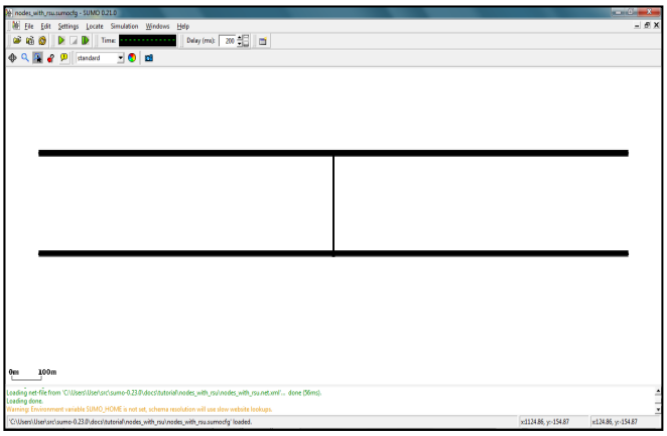


Figure 7: 2000m×1000m rectangular traffic network with ten junctions and 3×3 lane road in SUMO

Figure8: OMNET++ view of the road network with RSUs

Table 1: Simulation Parameters

Parameter	Value
Number of lanes	3+3
Number of scenario	5
Number of vehicles	22, 41, 61, 83 and 102
TxPower	20mW
TxDatasize	10Mbps
Number of slots per frame	256
Slot time	50ms
Registration time	2s
Simulation time	2 frames, on 258s to 288s

For comparison, we have devised five scenarios with 22, 41, 61, 83, and 102 vehicles and measured the performance of the mentioned protocols in cases where 10%, 15% and 20% of the registered vehicles were not using their assigned slots. Configuration of the simulation is listed in Table 1.

6.2. Simulation Result

In this section, we show performance comparison of DRWA-MAC with the aforementioned protocols.

Figure 9 shows performance comparison in terms of successfully transmitted data packets when 10% of the vehicles are not using their assigned slots. As shown in Figure 9, initially at light traffic, only FAWAC-MAC performed better than DRWA-MAC. However, the performance of FAWAC-MAC gradually decreased with the increase of traffic. In high traffic scenarios, the performance of VeMAC, ACFM, PTMAC and FAWAT-MAC degraded noticeably. This is because these protocols suffer from merging collision and one-hop neighboring collision. On the other hand, DRWA-MAC minimized these collisions and outperformed all the other schemes when network density is higher.

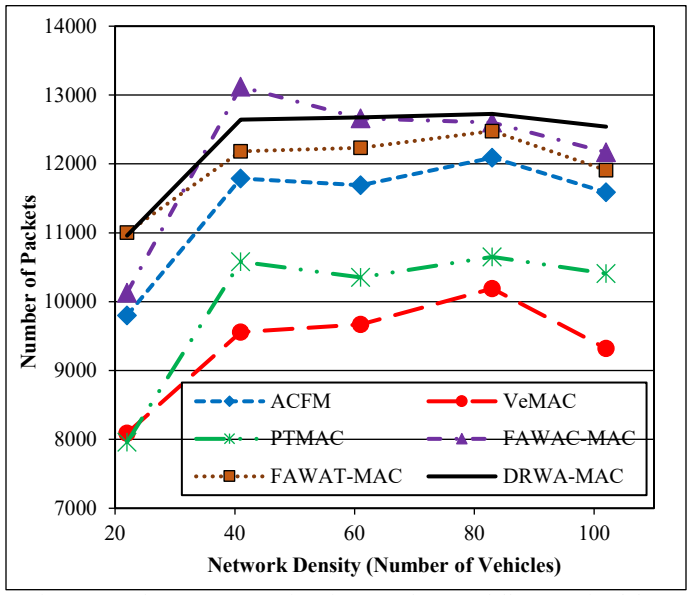


Figure 9: Performance comparison in term of successfully transmitted data packets when 10% of the vehicles are not using their assigned slots

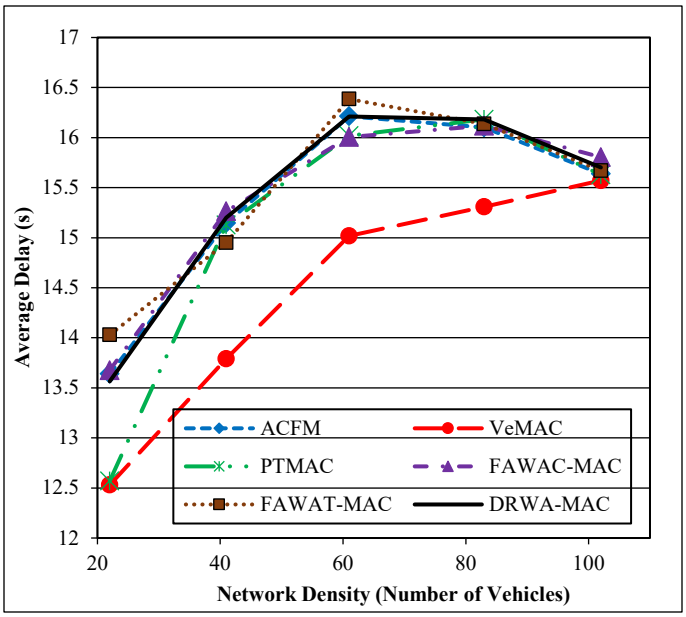


Figure10: Performance comparison in terms of average delay when 10% of the vehicles are not using their assigned slots

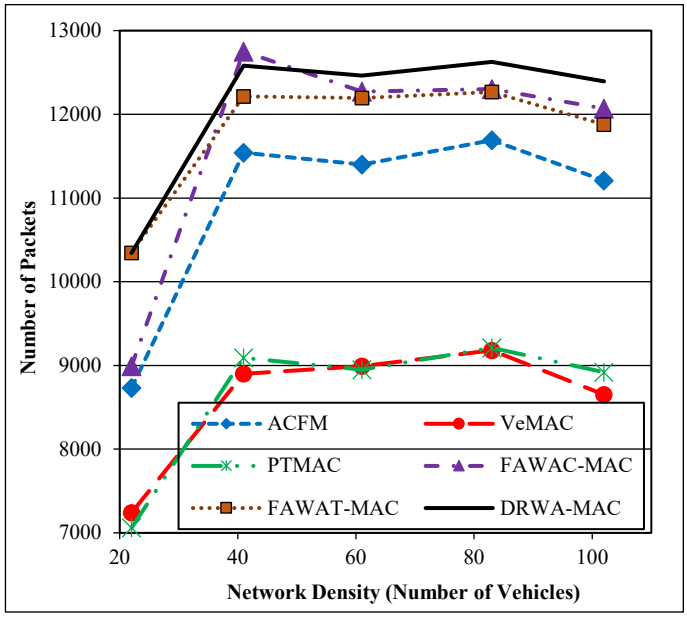


Figure 12: Performance comparison in terms of successfully transmitted data packets when 15% of the vehicles are not using their assigned slots

Figure 10 shows performance comparison in terms of average delay when 10% of the vehicles are not using their assigned slots. Though for a single channel, the overall throughput of VeMAC is not optimum, it has the most favorable average delay. On the other hand, our proposed DRWA-MAC performed better by maintaining almost the same average delay as ACFM, PTMAC, FAWAC-MAC and FAWAT-MAC.

Figure 11 shows performance comparison in terms of data loss ratio where 10% of the registered vehicles are not transmitting data in the assigned slots. Because of inherent huge low power transmission, as shown in Figure 11, FAWAC-MAC is having high data loss, though those low power packets were bearing opportunistic data. VeMAC has the lowest data loss ratio while DRWA-MAC has similar data loss ratio compared to other protocols.

In Figure 12 and 13, we show performance comparisons in terms of successfully transmitted data packets for cases where 15% and 20% of the registered vehicles are not transmitting data in their assigned slots respectively. With the increase of free slots, the performance of VeMAC, ACFM and PTMAC decreased as these protocols do not have any process to ensure reusability of slots. However, as FAWAC-MAC, FAWAT-MAC and DRWA-MAC

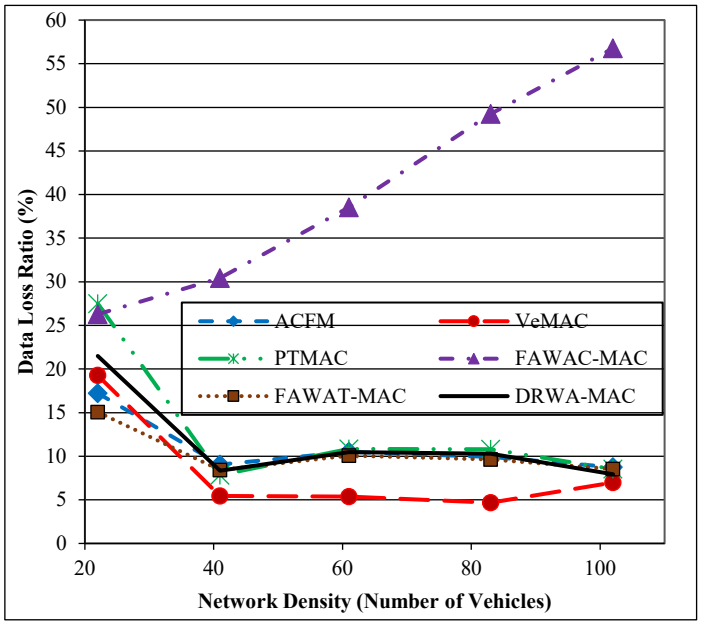


Figure11: Performance comparison in terms of data loss ratio when 10% of the vehicles are not using the slots assigned to them

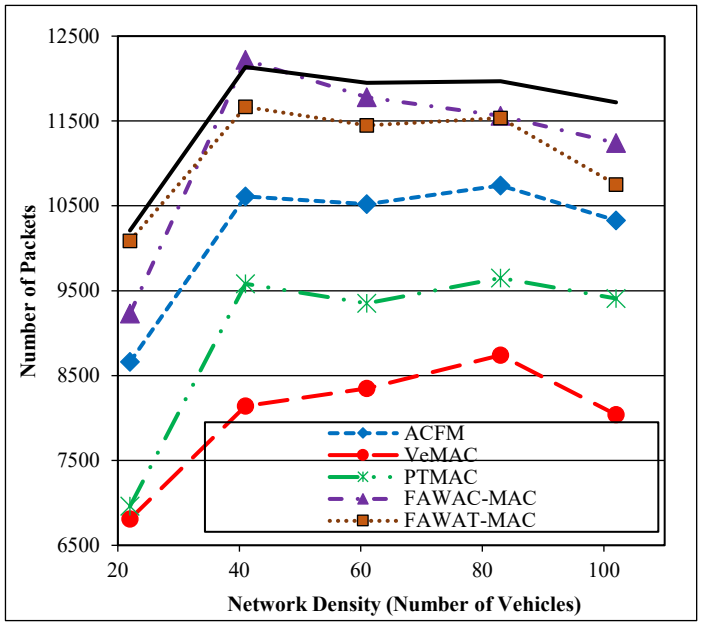


Figure 13: Performance comparison in terms of successfully transmitted data packets when 20% of the vehicles are not using their assigned slots

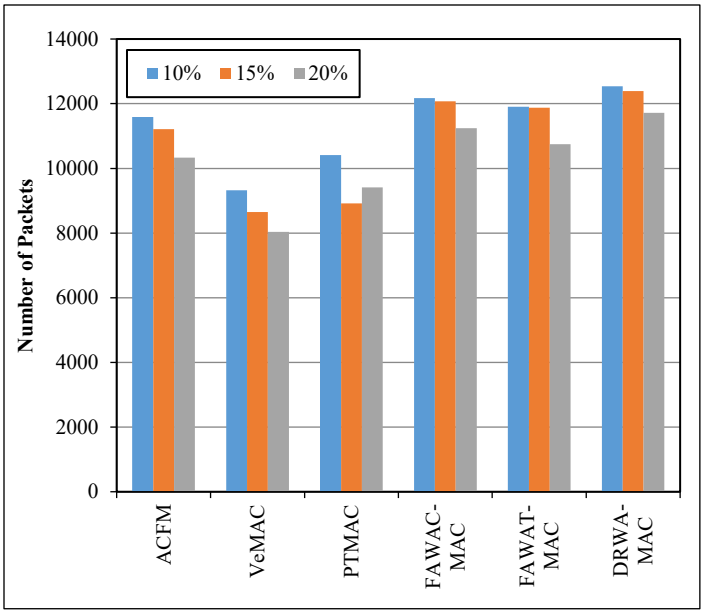


Figure 14: Bar diagram of successfully transmitted data packets by grouping scenarios where 10%, 15% and 20% of the vehicles are not using their assigned slots

can unsure reusability of free slots, their performance were not affected significantly. From Figure 9, 12 and 13, we can conclude that though FAWAT-MAC and DRWA-MAC improves the amount of successful data transmission by ensuring reallocation of free slots even though one fourth of the data capacity of those free slots degraded due to the detection delay of the RSU.

Figure 14 shows a bar diagram of successfully transmitted data packets for 102 cars in the network by grouping scenarios where 10%, 15% and 20% of the vehicles are not using their assigned slots. This diagram is also shows that with the increase of free slots, the performance of VeMAC, ACFM and PTMAC decreased, but the performance of FAWAC-MAC, FAWAT-MAC and DRWA-MAC were not affected significantly. As shown in the figure, DRWA-MAC outperforms all other schemes in terms of successfully transmitted data packet in all the experimental scenarios.

7. Conclusion

This article proposes a dynamic reallocation based window access scheme, DRWA-MAC, for centralized TDMA based VANETs. The proposed scheme maintains a balanced waiting time for vehicles by dividing TDMA slots into ODD slots and EVEN slots and then, by evenly distributing all TDMA slots among the registered vehicles in a rational way. Our slot allocation algorithm ensures fairness in slot allocation among the vehicles as well as reduces merging collision and one-hop neighboring collision. By dynamically monitoring and reallocating unused slots using “time slot reassignment” mechanism, the proposed DRWA-MAC achieved significantly higher throughput compared to other conventional TDMA based centralized protocols. Simulation results show the effectiveness of the proposed protocol in terms of successfully transmitted data packets, average delay and data loss ratio on overall road traffic network in different scenarios.

References

[1] Md. Amirul Islam, Shafika Showkat Moni and Mohammad Shah Alam, “A Fixed Assignment Based Window Access Scheme for Enhancing QoS of Vehicular Adhoc Networks” in International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017.

[2] Mohamed Hadded, Paul Muhlethaler, Anis Laouiti, Rachid Zagrouba, Leila Azouz Saidane, "TDMA-based MAC Protocols for Vehicular Ad Hoc Networks A Survey, Qualitative Analysis and Open Research Issues" inIEEE Communication Surveys and Tutorials, Volume: 17, Issue: 4, February 2015.

[3] Laouiti, Anis, Qayyum, Amir, Mohamad Saad, Mohamad Naufal, "Vehicular Ad-hoc Networks for Smart Cities" in First International Workshop(Advances in Intelligent Systems and Computing), 2014.

[4] M.J. Booysen, S. Zeadally, G.-J. Van Rooyen, "Survey of media access control protocols for vehicular ad hoc networks"in IET Communication, 2011, Vol. 5, Iss. 11, pp. 1619–1631.

[5] Georgios Karagiannis, Onur Altintas, Eylem Ekici, Geert Heijenk, Boangoat Jarupan, Kenneth Lin, and Timothy Weil, "Vehicular Networking: A Survey and Tutorial on Requirements, Architectures, Challenges, Standards and Solutions" in IEEE Communications Surveys & Tutorials, Volume: 13, Issue: 4, 2011.

[6] Mohammad S. Almalag, Stephan Olariu, Michele C. Weigle, "TDMA Cluster-based MAC for VANETs (TC-MAC)" in IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), June 2012.

[7] H. Omar, W. Zhuang, and L. Li, “VeMAC: A Novel Multichannel MAC Protocol for Vehicular Ad Hoc Networks” in IEEE Conf. Computer Communications Workshops (INFOCOM WORKSHOPS), April 2011, pp. 413–418.

[8] H. Omar, W. Zhuang, and L. Li, “VeMAC: A TDMA-Based MAC Protocol for Reliable Broadcast in VANETs” in IEEE Transactions on Mobile Computing, Volume: 12, Issue: 9, September 2013, pp. 1724 - 1736

[9] Weijie Guo, Liusheng Huang, Long Chen, Hongli Xu, Jietao Xie, "An Adaptive Collision-Free MAC Protocol Based on TDMA for Inter-Vehicular Communication" in International Conference on Wireless Communications & Signal Processing (WCSP), 2012.

[10] Xiaoxiao Jiang and David H.C. Du, "PTMAC: A Prediction-based TDMA MAC Protocol for Reducing Packet Collisions in VANET" in IEEE Transactions on Vehicular Technology (Volume: PP, Issue: 99), 19 January 2016.

[11] Christoph Sommer, “Veins - The open source vehicular network simulation framework”, Web Link “http://veins.car2x.org/”.

[12] András Varga, “OMNeT++, Discrete Event Simulator”, Web Link “https://omnetpp.org/”.

[13] DLR - Institute of Transportation Systems, “SUMO, Simulation of Urban MObility”, Web Link “http://sumo-sim.org/”.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 329-341 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Mission Profile Analysis of a SiC Hybrid Module for Automotive Traction Inverters and its Experimental Power-loss Validation with Electrical and Calorimetric Methods

Ajay Poonjal Pai [*1], Tomas Reiter[1], Oleg Vodyakho[2], Martin Maerz[3]

[1]*Infineon Technologies AG, Neubiberg, 85579, Germany*

[2]*Infineon Technologies Americas Corp, El Segundo, 90245, USA*

[3]*Fraunhofer IISB, Erlangen, 91058, Germany*

ARTICLEINFO

Article history:
Received: 15 November, 2017
Accepted: 09 January, 2018
Online: 02 February, 2018

Keywords:
Silicon Carbide
Automotive Traction Inverter
Calorimetric Measurement
Mission Profile
Power Loss Calculations
Power Analyser

ABSTRACT

This paper investigates the efficiency benefits of replacing the Silicon diodes of a commercial IGBT module for the main inverter application of an electric vehicle with Silicon Carbide diodes, leaving the package, operating conditions and the system unchanged. This ensures that the comparison is directly between the chip technologies without any scope for discrepancies arising out of differences in the packaging, gate-driver circuit etc. A behavioral power loss calculation model is used to investigate the performance of the two modules for various drive cycles (Artemis, WLTP, NEDC). The behavioral power loss model is experimentally validated using two independent measurement methods, namely, power analyser based electrical input output method, and a calorimetric method which was developed especially for the low lossy light load condition. Furthermore, it is shown that the electrical method has close to 30% inaccuracy making it unsuitable for the main inverter applications, especially for comparing two different chip technologies, e.g., Silicon versus Silicon Carbide. The developed calorimetric method in contrast offers lower than 3% uncertainty.

1 Introduction

This paper investigates the efficiency benefits of replacing a Si IGBT based power module of an automotive traction inverter with a SiC Hybrid module, for public mission profiles such as NEDC, WLTP and Artemis. This paper is an extension of the work presented in [1] and [2], where the hybrid-Silicon Carbide (SiC) module was characterized and mission profile analysis was performed for public mission profiles. In this work, additionally, the inverter power loss model used for analysis will be experimentally validated using two independent methods, firstly with the input-output based electrical method, and secondly with the calorimetric method presented in [3], suitable for automotive main inverters which operate at light load conditions, i.e., less than a quarter of the nominal current more than 90% of the time.

2 Review of Literature and Motivation

A vast number of papers have been published in the last two decades investigating the advantages of SiC in automotive converters, especially by car makers like Toyota [4] and Ford [5] among others. While a vast majority of the publications have been on dc-dc applications with high switching frequency, one can find only a few papers addressing the traction inverter application which is generally a low switching frequency application (8 to 20 kHz). These papers that address the topic of SiC for automotive traction inverters often have one or more of the following drawbacks:

1. The considered switching frequencies are far higher than typical application requirements, e.g., [6] presents a full-SiC automotive inverter, but the switching frequency considered

[*]Ajay Poonjal Pai, Infineon Technologies AG, Am Campeon 1-12, Neubiberg, Germany 85579 & AjayPoonjal.Pai@Infineon.com

https://dx.doi.org/10.25046/aj030140

is 50 kHz.

2. The considered devices are rated for low currents, in the range of 4 to 20 A, e.g., [7, 8, 9]. The resulting inverters are quite far from the typical application requirements of the traction inverter (above 100 A).

3. An impractical number of smaller devices are considered connected in parallel to meet the power ratings of a high power Insulated Gate Bipolar Transistor (IGBT) module. For example, [10] provides a comparison of SiC Junction gate Field Effect Transistors (JFETs) against Silicon (Si) IGBTs in traction inverter application for typical mission profiles. But in order to compare the 5 A SiC JFETs against a 300 A Si IGBT module, it is assumed that 60 JFETs are connected in parallel, which makes it an impractical and unfair comparison!

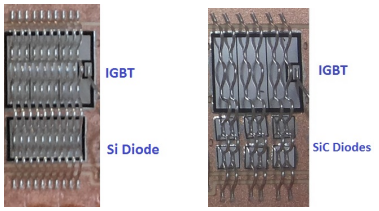
4. The devices and/or packages chosen for comparison are not suitable for mass production, but merely design studies, e.g., [11]. The constraints for a mass produced module can be quite different than for a design study module.

5. The compared Si and SiC chips are in completely different packages or application conditions, e.g., [12, 13, 14]. This makes it difficult to evaluate if the reported benefits of SiC are really coming from the advantages offered by the technology itself or simply due to the difference in package/operating conditions.

In short, a clear investigation of the efficiency benefits of using SiC as a direct replacement for a commercial Si- IGBT module at various boundary conditions, without giving any scope for discrepancies arising out of differences in the package, system, gate-driver circuit etc., is still missing in literature. This paper investigates the benefits of replacing a commercial Si IGBT module with a prototype Hybrid SiC module, as a first step, leaving the other components of the package and the system unchanged.

3 The Compared Modules

Infineon HybridPACK Drive [15] FS820R08A6P2B is chosen as the Si-IGBT module owing to its best-in-class low stray inductance (L_{sCE}=8nH) which makes it suitable for very fast switching applications. FS820R08A6P2 is an automotive qualified B6 bridge power module based on the new EDT2 Micro-pattern Trench-Field-Stop technology, with an implemented current rating of 820A per phase, and a blocking voltage rating of 750V. It has three IGBTs of 100 mm^2 each in parallel per switch (in total, A_I = 300 mm^2 per switch) and three anti-parallel diodes of 50 mm^2 each in parallel per diode (in total, A_D = 150 mm^2 per switch). Figure 1(a) shows an IGBT-diode pair of one switch. It comes with a Pin-Fin baseplate which is suitable for direct water cooling, and can operate up to T_j=175°C. It has one NTC per-phase integrated directly on the DCB which can be used for temperature sensing. For brevity, this module shall be referred to simply as "HPD" in this paper.



(a) With Si diode (b) With SiC diodes

Figure 1: An IGBT-diode pair of HybridPACK Drive Module

For the purpose of evaluating the benefits of SiC, a prototype SiC Hybrid module has been produced by replacing the Si diodes of HPD with Generation-5 650V SiC Schottky diodes from Infineon [16]. The SiC diodes have a die size of 7.12 mm^2 and a nominal current rating of 40A. Each of the 50 mm^2 Si diodes are replaced by a parallel connection of six SiC diodes as seen from figure 1(b), leaving the remaining construction of the module as it is, for a direct evaluation of the SiC diodes vis-a-vis the Si diodes. This module shall be referred to as "HPD-Hyb-SiC" in this paper.

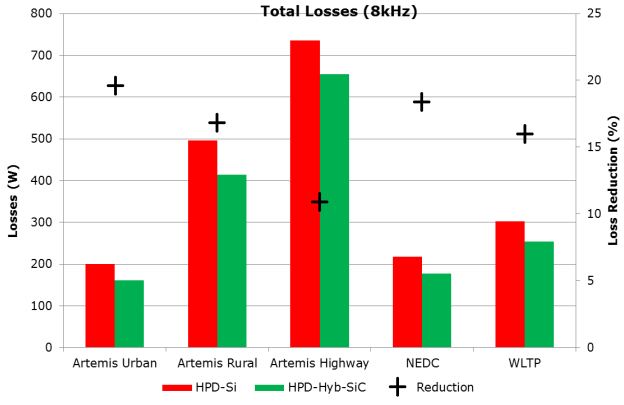
3.1 Inverter Power Loss Calculation Model

For a good comparison of the power loss performance of different chip technologies, it is imperative to have an accurate inverter power loss model. This model should offer an uncertainty significantly lower than the performance difference between the technologies compared. For example, it is meaningless to compare two different technologies which differ in loss performance by 20% with a model which suffers from 15% uncertainty, because it is not possible to ascertain if the apparent difference between the technologies is indeed because the technology is better than the other, or if it is merely due to the high uncertainty of the model used. Since this paper would compare Si and SiC technologies which usually differ by 10 to 20 percent, a model with an uncertainty of < ±5% presented in [17, 18] is used and the parameters A_{11}- E_{43} are determined for both the modules. The parameters for HPD can be found in [17, 18], and those for HPD-Hyb-SiC in [2].

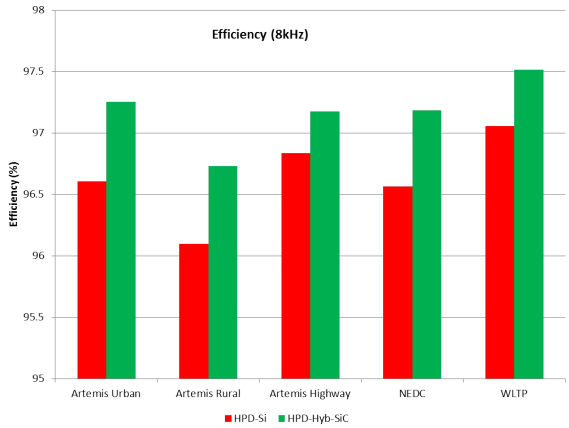
4 Mission Profile Analysis

The model discussed in the previous section is used to compare the performance of the two modules for 5 different mission profiles, viz., Artemis-Urban, Artemis-Highway, Artemis-Rural, WLTP and NEDC.

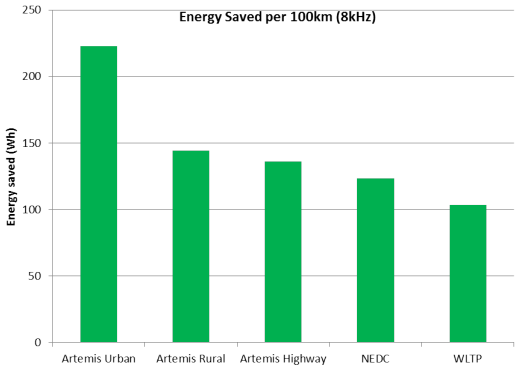
For the mission profile investigation, a typical mid size sedan similar to Volkswagen Golf presented in [2] is chosen as the reference vehicle. The results at 8kHz are given in figure 2.



(a)



(b)



(c)

Figure 2: Results of Mission Profile Analysis for the modules

As a result of the higher average driving speed, the overall losses are the highest for Artemis Highway driving Cycle. Artemis Urban on the contrary, has the least average losses due to its low average speed. HPD-Hyb-SiC has about 10-20% lower losses compared to HPD-Si. The reduction in the losses is the highest for the Artemis Urban cycle, where the inverter operates in the low current regime. This is because, at low currents the switching losses (which are significantly lesser in the SiC diode compared to the Si diode) dominate over the conduction losses (which are higher in the SiC diodes than the Si diodes). Next, the study is repeated at different switching frequencies, 8-15kHz. The improvement in efficiency of the SiC module over the Si module is summarized in figure 3. Again, the highest benefit of increasing f_{sw} is for the Artemis Urban drive cycle, which sees more than 1.2% improvement in efficiency at 15kHz compared to about 0.7% at 8kHz. Artemis Rural and Artemis Highway cycles too are not far behind. This is because in all the artemis cycles, the switching losses dominate over the conduction losses, and the scope for improvement with SiC diodes is high. The NEDC and WLTP cycles are mostly dominated by conduction losses, and as a result the benefit of using SiC diodes is not high.

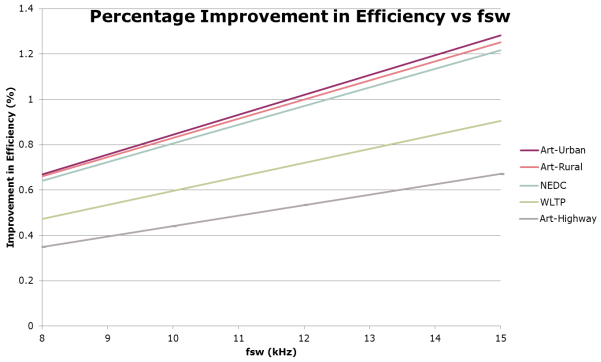


Figure 3: Mission Profile analysis at different f_{sw}

5 Experimental Validation of the Behavioral Power Loss Calculation Model

To have a good confidence level in the behavioral power loss calculation model used for mission profile analysis and to verify it experimentally, it is important to measure the inverter power losses at various operating points, preferably with two independent methods. Firstly, power loss measurements are performed at various inverter operating points with the power-analyser based input-output electrical method, which, being relatively easier to perform, is the most commonly used method for such applications. The sources of uncertainty are thoroughly analysed and it will be shown that this method has high uncertainty due to the switched nature of the output voltage waveform. A common solution to this problem is to add a sine output filter. However, the results include also the losses in the filter leading to wrong results, as will be shown in this paper. Thus, it is necessary to resort to the calorimetric method proposed in [3], which is nearly as easy to perform as the electrical method and is particularly suitable for low-lossy conditions.

This method does not require the use of an expensive calorimeter.

6 The Electrical Input-Output-based Method

6.1 Test Setup

A test platform is built as per the schematic in figure 4. The inverter (described in section 7) is connected to a 3-phase inductive load. The input dc current I_{dc} and output rms currents I_{rms} are sensed using high accuracy closed loop current transducers IT 200-S ULTRASTAB [19] and LF 510-S [20] respectively, which are then fed into a state-of-the-art precision Power Analyser WT1800 [21]. WT1800 has a high sampling frequency of 2 MHz with 16-bit resolution and a bandwidth of 5 MHz. It is also equipped with a digital line filter which can be set for frequencies from 100 Hz to 100 kHz. The impact of using this line filter on the accuracy will be also discussed later in this paper. The inverter is run in open-loop mode, and the desired output voltage is requested by setting the modulation index accordingly. The power analyser measures the input dc power P_{dc} and the output ac power P_{ac}, and the difference is equal to the power loss P_{loss}, from the basic definition, as per equation 1.

$$P_{loss} = P_{dc} - P_{ac} \tag{1}$$

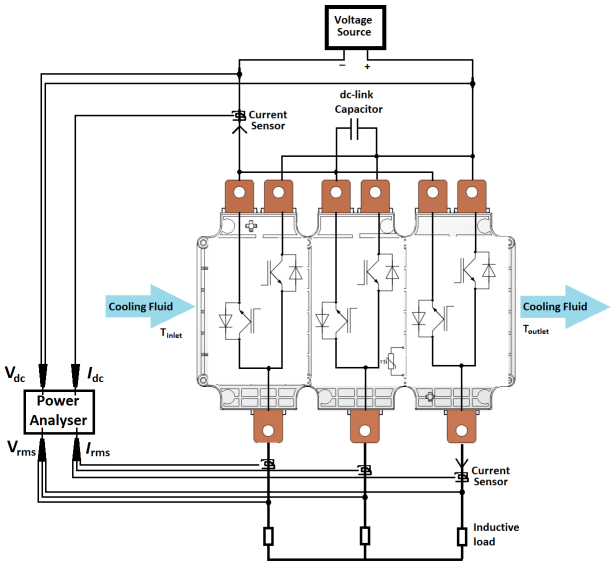


Figure 4: Schematic for measurement of power losses with the electrical method

6.2 Sources of Uncertainty

Uncertainties occur in this method, on account of the delays introduced by the probes and unintended phase-shifts between the different measured signals. In this section, the uncertainty involved in this method will be systematically derived. Out of the scope of this work are the errors introduced on account of the Radio Frequency Interference (RFI) and Electro Magnetic Interference (EMI) emanating from the high di/dt and dv/dt prevailing in hard switched power converters.

The uncertainty in the measurement of the power loss ΔP_{loss} from equation 1 can be derived using the Gaussian law of error propagation.

$$\frac{\Delta P_{loss}}{P_{loss}} = \sqrt{\left(\frac{\Delta P_{dc}}{P_{dc}}\right)^2 + \left(\frac{\Delta P_{ac}}{P_{ac}}\right)^2} \tag{2}$$

Equation 2 gives the uncertainty with a confidence value of about 68%. This will simply be referred to as *uncertainty* in the rest of this work. The upper bound for the uncertainty $\Delta P_{loss,max}$ can be simply calculated as

$$\frac{\Delta P_{loss,max}}{P_{loss}} = \frac{\Delta P_{dc}}{P_{dc}} + \frac{\Delta P_{ac}}{P_{ac}} \tag{3}$$

This will hence be referred to as the *maximum uncertainty*.

6.2.1 Uncertainty in P_{dc}

To calculate ΔP_{dc}, we have to go back to the fundamental equation of dc power, i.e.,

$$P_{dc} = V_{dc} \cdot I_{dc} \tag{4}$$

and the uncertainty can again be calculated by the Gaussian law described above as:

$$\frac{\Delta P_{dc}}{P_{dc}} = \sqrt{\left(\frac{\Delta V_{dc}}{V_{dc}}\right)^2 + \left(\frac{\Delta I_{dc}}{I_{dc}}\right)^2} \tag{5}$$

and the maximum uncertainty as:

$$\frac{\Delta P_{dc,max}}{P_{dc}} = \frac{\Delta V_{dc}}{V_{dc}} + \frac{\Delta I_{dc}}{I_{dc}} \tag{6}$$

From the reference manual of the power analyser [21], the uncertainties for the dc voltage measurement are given as a function of the set range $V_{dc,range}$ and the reading itself as follows

$$\Delta V_{dc} = 0.0005 \cdot V_{dc} + 0.001 \cdot V_{dc,range} \quad [V] \tag{7}$$

Additionally, it must be remembered that there is also a ripple in V_{dc} consisting mainly of a second harmonic of the fundamental frequency, and small magnitudes of harmonics of the switching frequency. These harmonics also contribute to the uncertainty, as will be covered in section 6.2.2. But, as the magnitude of this ripple is quite small, their contribution to the uncertainty can be neglected.

The uncertainties for the measurement of I_{dc} are similarly given as

$$\Delta I_{dc,analyser} = 0.0005 \cdot I_{dc} + 0.001 \cdot I_{dc,range} \quad [A] \tag{8}$$

$\Delta I_{dc,analyser}$, however, is the uncertainty of just the power analyser and since we are using an external LEM current transducer, we also have to take its uncertainty into account. From [19], the uncertainty

for the LEM transducer on the secondary side can be given as:

$$\Delta I_{\mathrm{dc,LEM,max,sec}} = I_{\mathrm{OE}} + \epsilon_{\mathrm{L}} \cdot I_{\mathrm{pry,range}} \cdot K_{\mathrm{N}} \quad [A] \quad (9)$$

and

$$\Delta I_{\mathrm{dc,LEM,sec}} = \sqrt{I_{\mathrm{OE}}^2 + (\epsilon_{\mathrm{L}} \cdot I_{\mathrm{pry,range}} \cdot K_{\mathrm{N}})^2} \quad (10)$$

where, $I_{\mathrm{OE}} = 80 \cdot 10^{-6}$A is the electrical offset current, $\epsilon_{\mathrm{L}} = 3 \cdot 10^{-6}$A is the linearity error, $I_{\mathrm{pry,range}}$=200A is range on the primary side and K_{N}=0.001 is the turns ratio. It is to be noted that, the effects due to self heating in the current transducer are neglected and it is assumed that the transducer is at a constant temperature of 25 °C. Since we are interested in the primary current for calculating P_{dc}, the absolute uncertainties get divided by the turns ratio as follows:

$$\Delta I_{\mathrm{dc,LEM}} = \frac{\Delta I_{\mathrm{dc,LEM,sec}}}{K_{\mathrm{N}}} \quad (11)$$

$$\Delta I_{\mathrm{dc,LEM,max}} = \frac{\Delta I_{\mathrm{dc,LEM,max}}}{K_{\mathrm{N}}} \quad (12)$$

Clubbing equations 8, 9 and 11/12, we obtain the effective uncertainty for I_{dc} as:

$$\frac{\Delta I_{\mathrm{dc}}}{I_{\mathrm{dc}}} = \sqrt{\left(\frac{\Delta I_{\mathrm{dc,analyser}}}{I_{\mathrm{dc,analyser}}}\right)^2 + \left(\frac{\Delta I_{\mathrm{dc,LEM}}}{I_{\mathrm{dc,LEM}}}\right)^2} \quad (13)$$

$$\frac{\Delta I_{\mathrm{dc,max}}}{I_{\mathrm{dc}}} = \frac{\Delta I_{\mathrm{dc,analyser}}}{I_{\mathrm{dc,analyser}}} + \frac{\Delta I_{\mathrm{dc,LEM,max}}}{I_{\mathrm{dc,LEM,max}}} \quad (14)$$

6.2.2 Uncertainty in P_{ac}

From the fundamental equation for the ac power for a line-line voltage of V_{rms}, line current I_{rms} and a power factor angle ϕ (in radians),

$$P_{\mathrm{ac}} = \sqrt{3} \cdot V_{\mathrm{rms}} \cdot I_{\mathrm{rms}} \cdot \cos\phi \quad (15)$$

and the uncertainty can again be calculated by the Gaussian law described above as:

$$\frac{\Delta P_{\mathrm{ac}}}{P_{\mathrm{ac}}} = \sqrt{\left(\frac{\Delta V_{\mathrm{rms}}}{V_{\mathrm{rms}}}\right)^2 + \left(\frac{\Delta I_{\mathrm{rms}}}{I_{\mathrm{rms}}}\right)^2 + (\Delta\phi \cdot \tan\phi)^2} \quad (16)$$

When ϕ is expressed in degrees, we have:

$$\frac{\Delta P_{\mathrm{ac}}}{P_{\mathrm{ac}}} = \sqrt{\left(\frac{\Delta V_{\mathrm{rms}}}{V_{\mathrm{rms}}}\right)^2 + \left(\frac{\Delta I_{\mathrm{rms}}}{I_{\mathrm{rms}}}\right)^2 + \left(\frac{\pi}{180} \cdot \Delta\phi \cdot \tan\phi\right)^2} \quad (17)$$

The calculation of uncertainty in the measurement of P_{ac} is less straightforward than that for P_{dc}. This is because, the inverter output voltages and currents used for determining P_{ac} are not pure sinusoids and contain a spectrum of various frequencies, and the uncertainties of the power analyser are differently defined for the spectral components of different frequencies. Moreover, the reference manuals for the power analyser define the uncertainties only for pure sine waves and do not explicitly describe how these spectral components have to be treated. One simplified approach, commonly adopted in literature, is to assume that the output voltage spectrum would be dominated by the fundamental output frequency f_{o} and then calculate the uncertainty as if the output voltage were a pure sine wave with f_{o}. However, as can be expected and as shall be shown later in this paper, such an approach results in a significant discrepancy in the estimation of uncertainty. Therefore, in this work, we propose the following spectrum-based approach.

Spectrum-based Approach to Calculating Uncertainty of Non-sinusoidal Signals

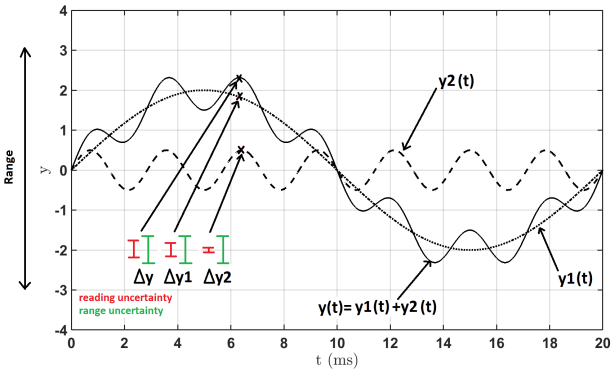


Figure 5: Reading- and range-uncertainties for a summation of signals

To understand how reading- and range- uncertainties should be calculated for a signal that is a combination of several signals, consider figure 5 which shows three signals, viz. $y_1(t)$ at 50 Hz, $y_2(t)$ at 350 Hz and $y(t)$= $y_1(t)$+$y_2(t)$. Lets assume that all the signals are measured with the same range r. Lets consider the points (t,y_1), (t,y_2) and (t,y) on the three signals respectively. The uncertainties for the three signals are also shown. The reading uncertainties $\Delta y_{1/2,\mathrm{reading}}$ can be calculated as follows:

$$\Delta y_{1,\mathrm{reading}} = \epsilon_{\mathrm{reading}}(f)\Big|_{f=50Hz} \cdot y_1 \quad (18)$$

$$\Delta y_{2,\mathrm{reading}} = \epsilon_{\mathrm{reading}}(f)\Big|_{f=350Hz} \cdot y_2 \quad (19)$$

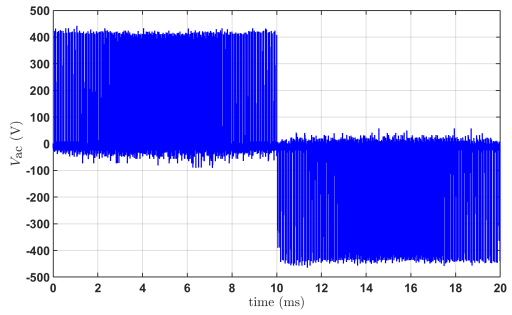
where, $\epsilon_{\mathrm{reading}}(f)$ is the frequency-dependent reading uncertainty coefficient normally specified in the reference manual. The range uncertainties $\Delta y_{1/2,\mathrm{range}}$ can be calculated as follows:

$$\Delta y_{1,\mathrm{range}} = \Delta y_{2,\mathrm{range}} = \epsilon_{\mathrm{range}} \cdot r \quad (20)$$

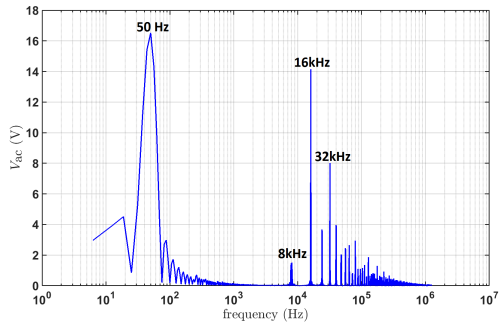
where, $\epsilon_{\mathrm{range}}$ is the uncertainty coefficient for the range, also specified in the reference manual. As the dependency of the range uncertainty on the spectral

components is small, it is assumed that ϵ_{range} is independent of f. The reading uncertainty $\Delta y_{\text{reading}}$ for the summed signal can be calculated simply by the sum of the reading uncertainties for the two signals as follows: The uncertainties for the three signals are also shown. The range uncertainties Δy_{range} can be calculated as follows:

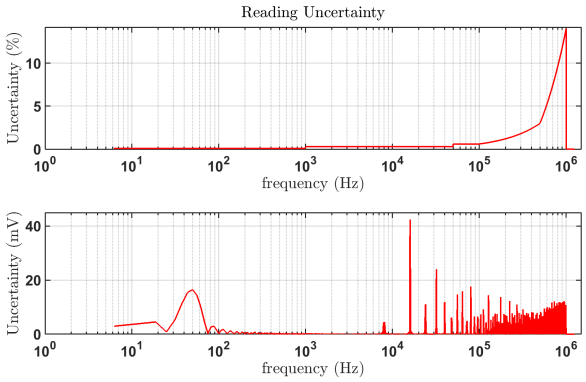
$$\Delta y_{\text{reading}} = \Delta y_{1,\text{reading}} + \Delta y_{2,\text{reading}} \tag{21}$$



(a) A fundamental period of V_{ac} in time domain



(b) Spectral components



(c) Reading uncertainty for the spectral components

Figure 6: Measured output line-line voltage V_{ac} of a hard-switched inverter

The range uncertainty Δy_{range}, however, could be expected to be the same for Δy as for $\Delta y_{1/2}$ as we are using the same range. This means that, unlike in the case of the reading uncertainty, the range uncertainty for the sum of two signals is not equal to the sum of the range uncertainties for the two signals. Therefore, Δy_{range} is equal to that calculated by equation 20.

Let us now consider a practical example. Figure 6(a) shows one fundamental period of the measured waveform of a typical hard-switched inverter output line-line voltage V_{ac}. First, the measured voltage waveform is subject to a fourier transformation to decompose it into its spectral components, as shown in figure 6(b). It can be seen that apart from the fundamental frequency 50Hz, there is significant contribution at the switching frequency 8kHz and multiples thereof. The reading uncertainties for the different components are calculated using the respective equations in table 1, as defined by the reference manual. Figure 6(c) shows the calculated reading uncertainty, both absolute and percentage, as a function of the spectral frequency. The original spectrum taking into account the uncertainties is now transformed back into the time domain. The range uncertainty is calculated as described previously and is added to this time-domain signal to yield the required $V_{\text{ac}} + \Delta V_{\text{ac}}$.

Table 1: Definition of Uncertainty for different Spectral Components

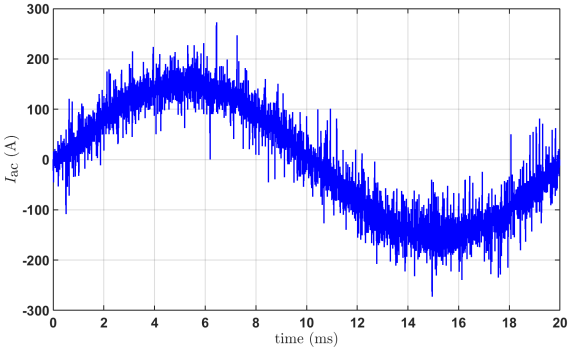
DC	$\Delta V_{\text{rms}} = 0.0005 \cdot V_{\text{rms}} + 0.001 \cdot V_{\text{rms,range}}$
0.1-10 Hz	$\Delta V_{\text{rms}} = 0.001 \cdot V_{\text{rms}} + 0.002 \cdot V_{\text{rms,range}}$
10-45 Hz	$\Delta V_{\text{rms}} = 0.001 \cdot V_{\text{rms}} + 0.001 \cdot V_{\text{rms,range}}$
45-66 Hz	$\Delta V_{\text{rms}} = 0.001 \cdot V_{\text{rms}} + 0.0005 \cdot V_{\text{rms,range}}$
66-1000 Hz	$\Delta V_{\text{rms}} = 0.001 \cdot V_{\text{rms}} + 0.001 \cdot V_{\text{rms,range}}$
1-50 kHz	$\Delta V_{\text{rms}} = 0.003 \cdot V_{\text{rms}} + 0.001 \cdot V_{\text{rms,range}}$
50-100 kHz	$\Delta V_{\text{rms}} = 0.006 \cdot V_{\text{rms}} + 0.002 \cdot V_{\text{rms,range}}$
100-500 kHz	$\Delta V_{\text{rms}} = (0.00006 * f) \cdot V_{\text{rms}} + 0.005 \cdot V_{\text{rms,range}}$
500-1000 kHz	$\Delta V_{\text{rms}} = (0.00022 * f - 0.08) \cdot V_{\text{rms}} + 0.01 \cdot V_{\text{rms,range}}$

This spectrum-based approach is also used to calculate the uncertainty for the current I_{ac}. The measured waveform is shown in figure 7(a). It can be seen that, unlike V_{ac}, I_{ac} is nearly sinusoidal which is on account of the inductance of the load. This can also be verified from the frequency spectrum shown in figure 7(b) where it can be seen that most of the energy is concentrated at the fundamental frequency. The contribution of the harmonics of the fundamental and the switching frequency is significantly lesser than in the case of V_{ac}. The calculated uncertainty as a function of the frequency spectrum is shown in figure 7(c).

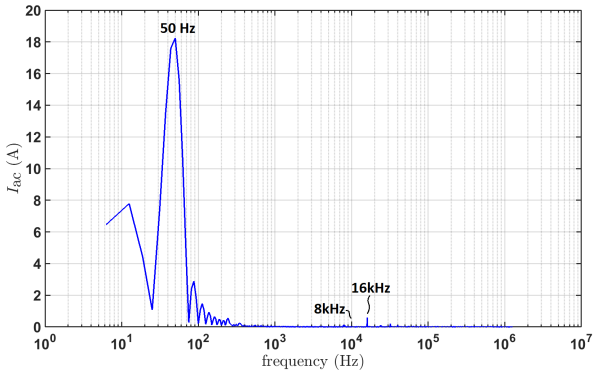
The total uncertainties thus calculated for the rms values of V_{ac} and I_{ac} are tabulated in table 2[1]. The approximate approach underestimates the uncertainty in V_{ac} by a factor of three compared to the spectrum-based approach, due to the abundance of high-frequency content in the waveforms. For I_{ac}, on the other hand, both approaches result in nearly the same value, owing to the current waveform being nearly sinusoidal. Therefore, it can be summarized that the approximated approach is sufficient for calculating uncertainty in I_{ac}, but it is necessary to go to

[1] As with I_{dc}, we are using an external LEM current transducer for measurement and therefore, we must separately calculate the uncertainties $\Delta I_{\text{rms,analyser}}$ and $I_{\text{rms,LEM}}$ for the power analyser and the LEM transducer [20] respectively.

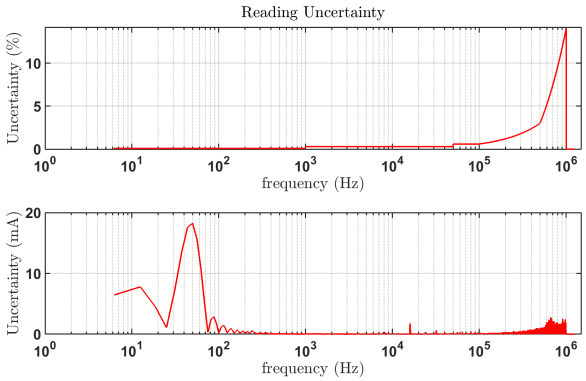
the more exact spectrum-based approach for calculating uncertainty in V_{ac}. Moreover, in applications with a higher switching frequency or a higher operating dc-link voltage V_{dc}, which is typical for a Silicon-Carbide-based application, the uncertainty in V_{ac} is higher respectively due to a wider distribution in the spectrum and a higher range that has to be chosen. In such applications, a higher deviation can be expected between the approximate and the spectrum-based approach which makes it more meaningful to use the spectrum-based approach.



(a) A fundamental period of I_{ac} in the time-domain



(b) Frequency spectrum



(c) Reading uncertainty for the spectral components

Figure 7: Measured line current I_{ac}

The uncertainty contribution due to the current transducer $\Delta I_{rms,LEM}$ can be calculated as follows:

$$\Delta I_{rms,LEM} = 0.005 \cdot I_{rms,nom} \tag{22}$$

where, the nominal current $I_{rms,nom}$=500A for the device used. The uncertainty ΔI_{rms} can be now written as

Table 2: Calculated measurement uncertainty for V_{rms} and I_{rms}

V_{rms}	ΔV_{rms}	
	Approximate	Spectrum-based
185.39 V	0.39 V (0.2 %)	1.11 V (0.6 %)
I_{rms}	$\Delta I_{rms,analyser}$	
	Approximate	Spectrum-based
106.158 A	0.181 A (0.17%)	0.203 A (0.19%)

$$\frac{\Delta I_{rms}}{I_{rms}} = \sqrt{\left(\frac{\Delta I_{rms,analyser}}{I_{rms,analyser}}\right)^2 + \left(\frac{\Delta I_{rms,LEM}}{I_{rms,LEM}}\right)^2} \tag{23}$$

and the maximum uncertainty as

$$\frac{\Delta I_{rms,max}}{I_{rms}} = \frac{\Delta I_{rms,analyser}}{I_{rms,analyser}} + \frac{\Delta I_{rms,LEM}}{I_{rms,LEM}} \tag{24}$$

For WT-1800, $\Delta\phi$ is given as

$$\Delta\phi = \left(\left|\phi - \cos^{-1}\left(\frac{\lambda}{1.002}\right)\right|\right) + \sin^{-1}\left(0.001 \cdot V_{rms} \cdot I_{rms}\right) \tag{25}$$

where λ is the power factor. Lastly, it has to be remembered that the uncertainty of the power analyser depreciates over the passage of time from its recent calibration. For WT-1800, the uncertainty at one year is 1.5 times that at 6 months.

6.2.3 A Common Measurement Mistake while applying Line Filters in Power Analysers

Most state-of-the-art power analysers come equipped with digital line filters which can be used to attenuate spectral components in the measured signals with a frequency higher than a certain cut-off frequency (generally programmable individually for each input channels). These filters are meant to be used on measurement signals where there is high frequency noise due to the limitation of the measuring equipment. Let us suppose that such a filter is used on the I_{ac} signal with a cut-off fequency of, say, 1kHz. The high frequency components in I_{ac} are predominantly due to measurement noise as can be seen from figure 7(b) and table 2, and using such a filter would help in attenuating this noise, thereby making the measurements more meaningful. However, suppose we use the same filter, either intentionally or accidentally, on V_{ac} which inherently has a high-frequency content (see figure 6(b) and table 2), other than measurement noise. In such a case, the filter would attenuate not only the noise, but also these high frequency components. This would, in turn, result in a lower-than-real measured value for P_{ac} and therefore, a higher P_{loss}. Therefore, the line filter should be used only on signals which do not inherently have high frequency components, like V_{dc}, I_{dc} and I_{ac}, but not on signals

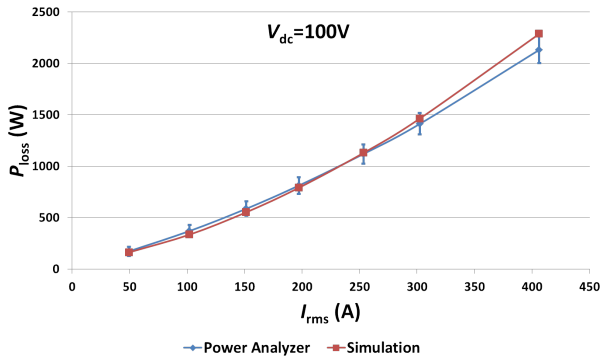
like V_{ac} which have a high frequency content. This is a common mistake while performing measurements incorporating such line filters, and will be demonstrated in the next section 6.3.

6.3 Measurement Results

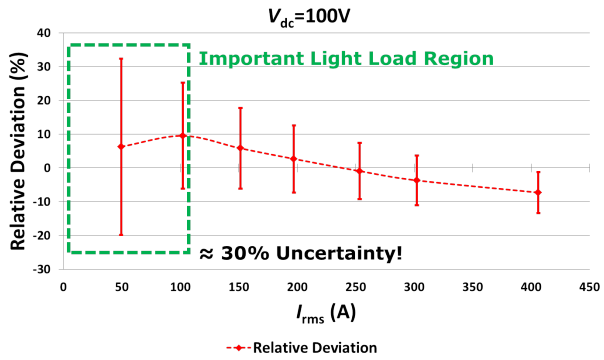
Figures 8-12 show the power losses measured with the electrical method for different application conditions. The uncertainties calculated for each of the measurements, using the spectrum-based approach described in the previous section, are shown as bars around the measurement points. Also shown are the simulated results based on the behavioral model discussed previously, and the relative deviation of the measured values from the simulation.

Figure 8(a) and 9(a) show P_{loss} measured for different output rms currents, I_{rms}, at V_{dc}=100V and 300V respectively. Across the entire range of the measured current, it can be seen that the simulations are within the tolerance of this state-of-the-art electrical input-output measurement approach, thereby validating the behavioral power loss calculation model.

In figures 10(a) and 11(a), the results are shown for different values of the gate resistances $R_{g,on}$ and switching frequencies f_{sw}. It can again be seen that the simulations are within the tolerance of this state-of-the-art input-output measurement approach.

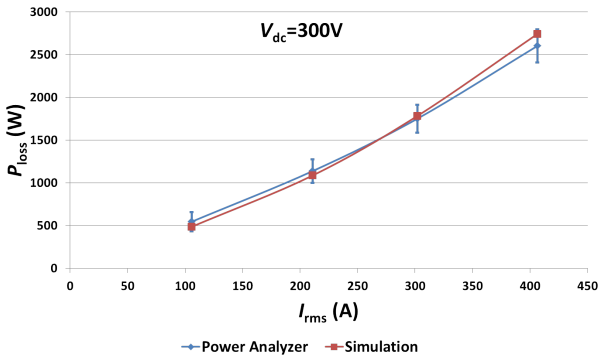


(a) P_{loss}

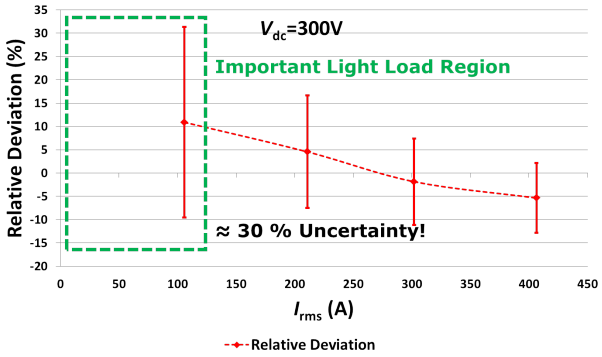


(b) Relative Deviation (%)

Figure 8: Comparison of the electrical method with simulations: P_{loss} vs. I_{rms} at V_{dc}=100V

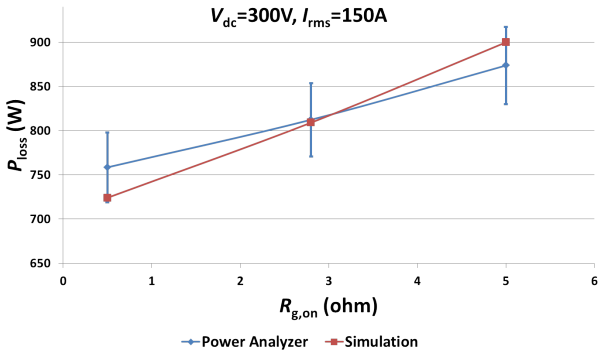


(a) P_{loss}

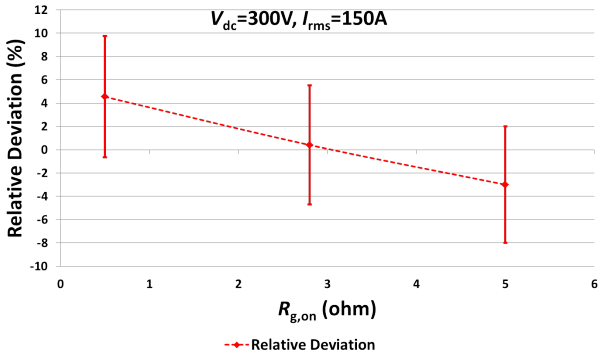


(b) Relative Deviation (%)

Figure 9: Comparison of the electrical method with simulations: P_{loss} vs. I_{rms} at V_{dc}=300V



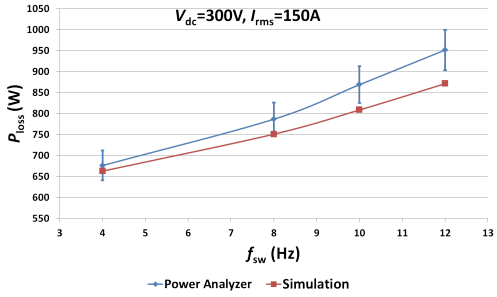
(a) P_{loss}



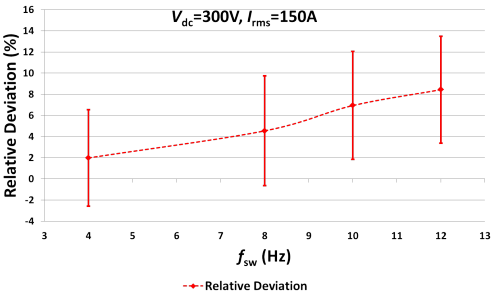
(b) Relative Deviation (%)

Figure 10: Comparison of the electrical method with simulations: P_{loss} vs. $R_{g,on}$

Lastly, figure 12 shows the measurements performed with the line filter (cut-off frequency=1kHz) enabled for the output ac voltage V_{ac}, and compares them with the measurements without the filter. As explained in section 6.2.3, it can be seen that the losses measured are higher without the line filter because the measured P_{ac} is lower-than-real. This shows that enabling the line filter on V_{ac} can lead to wrong measurements. Overall, it must be observed that the electrical input-output based method has nearly 30% uncertainty in the light-load condition, making this method unsuitable for the main inverter application, especially when comparing different chip generations.



(a) P_{loss}



(b) Relative Deviation (%)

Figure 11: Comparison of the electrical method with simulations: P_{loss} vs. f_{sw}

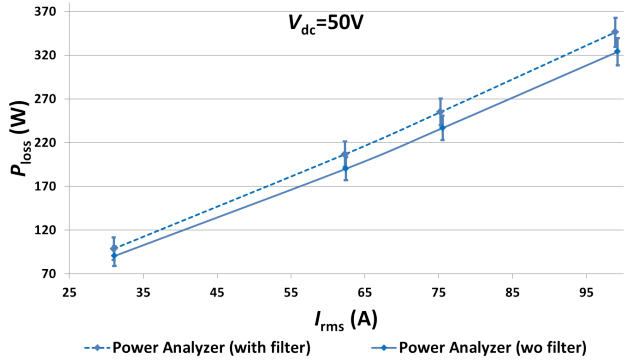


Figure 12: Comparison of the electrical method with simulations: Effect of Line Filter

7 Calorimetric Measurement of Power Losses

A more accurate approach, compared to the electrical method, is to measure power losses with a calorimeter [22]. For applications such as the automotive main inverter which operate at less than a quarter of the inverter nominal current more than 90% of the time[17], the light-load low-lossy condition is of interest. With traditional calorimetric methods, however, a sufficient rise in the fluid temperature is hard to obtain at these conditions. To overcome this problem, without compromising on the accuracy, the inverter is subjected to a calorimetric method particularly suitable for low-lossy conditions. This method does not require the use of an expensive calorimeter and is presented in detail in [3]. A requirement for the test is that all the energy losses in the module must ideally go into heating up the baseplate, with no convection. Therefore, the module baseplate is thermally insulated with a layer of polystyrene as can be seen in figure 13. EVAL-6ED100HPDRIVE-AS, a 6-channel gate-driver board based on the EiceDriver 1EDI2001AS from Infineon, is connected on the top of the module. The gate-driver board is controlled by a micro-controller logic board connected on top of it. The logic board is connected to a computer through a CAN bus, and the parameters such as f_{sw}, m, f_{out} can also be controlled by software, in open loop or closed loop modes. Moreover, it is also possible to read the temperatures sensed by the NTCs and log them. The complete inverter system is shown in figure 14. This method comprises the following two stages.

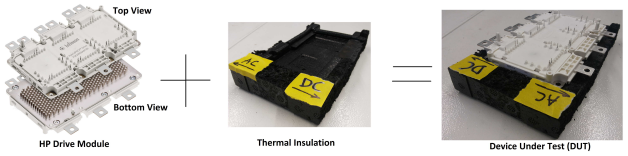


Figure 13: The DUT with the Thermal Package for the Calorimetric Test Bench

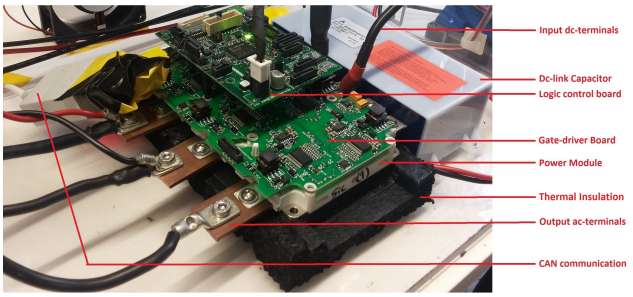
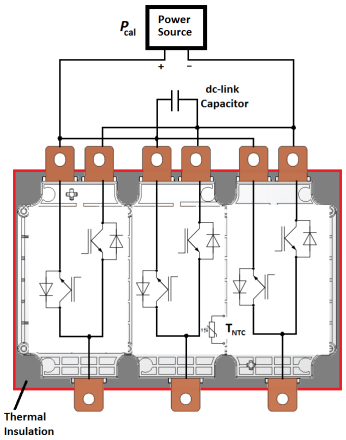


Figure 14: The Complete Inverter System in the Calorimetric Test Bench with the Thermally Isolated Baseplate

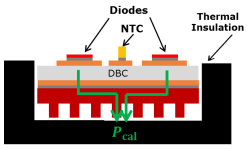
7.1 Calibration Stage

In this stage, the inverter system is connected to a dc source (with opposite polarity) with voltage sense pins capable of serving as a constant power source. The output ac terminals are disconnected as shown

in the schematic in figure 15(a). A known amount of constant power P_{cal} is injected into the module through the diodes. As there is no output power, all the injected power is dissipated in the module as heat, which is trapped in the heatsink on account of the thermal insulation, as seen in figure 15(b). As the temperature of the diode increases, its voltage drop changes and the dc source must be capable of suitably adjusting the current to maintain the power constant. Due to the thermal insulation, and the absence of convection, almost all the heat is trapped in the capacitance of the baseplate, and goes on to increase its temperature exponentially as shown in figure 16(a).



(a) Schematic Diagram



(b) Thermal Stack

Figure 15: Calibration Setup

There is a small amount of heat that is radiated into the ambient, or leaves the modules through any surface other than the heatsink. However, this effect will be cancelled out and have no impact on the accuracy of this method if the test setup during the calibration and measurement stages is identical. The temperature sensed by the NTCs is recorded until it rises from $T_1 = 50°C$ to $T_2 = 90°C$[2], after which the dc source is switched off and the system is let to cool down to the ambient temperature. In order to filter out measurement noise, an exponential curve is fitted to the measurement. From this fitted curve, the time taken for the temperature to reach T_1 from T_2 is taken as the *rise time* t_r. The temperature *slew rate* s_r is calculated as

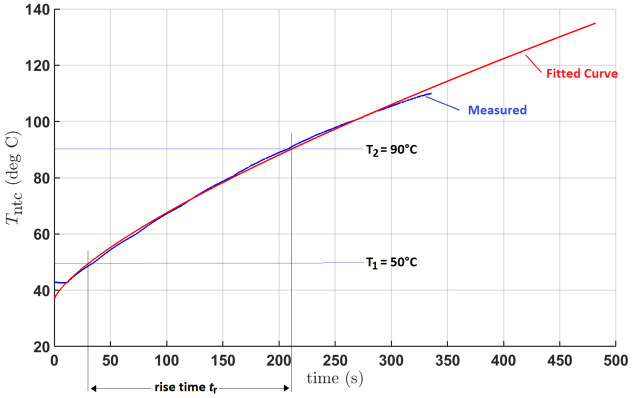
$$s_r = \frac{T_2 - T_1}{t_r} \tag{26}$$

This experiment is repeated at different values of injected power and P_{cal} is plotted against s_r as shown in figure 16(b) for both the modules and first order curves are now fitted to the two curves respectively, and the equations of the fitted curves are given below:

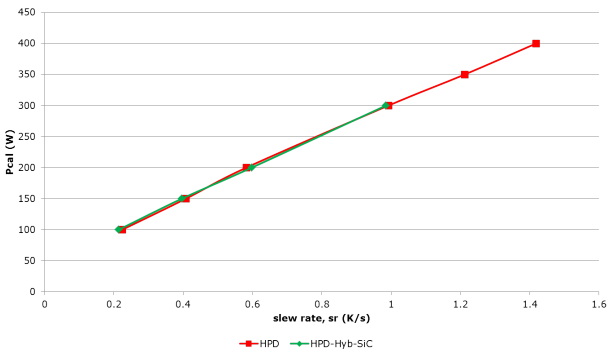
$$P_{loss} = 249.68 \cdot s_r + 48.655 \quad \text{[HPD Module]} \tag{27}$$

$$P_{loss} = 258.15 \cdot s_r + 45.905 \quad \text{[HPD-Hyb-SiC Module]} \tag{28}$$

As seen in figure 16(b), the calibration curves for the two modules match very closely, owing to the similar construction of the module. This is in line with the objective of this work to have minimum discrepancies arising out of differences in the packaging, for a fair comparison of Si and SiC. The slight difference seen between the curves can be attributed to the difference in the NTCs and the temperature measurement tolerances, which are cancelled out due to this method.



(a) NTC temperature at P_{cal} =100W



(b) Calibration Curve

Figure 16: Calibration

7.2 Measurement Stage

In this stage, the inverter is connected to a dc voltage-source (with normal polarity). Care must be taken to see that the setup, particularly the thermal insulation, is not disturbed between this stage and the calibration stage. The output ac terminals are connected to a three-phase star-connected passive load as shown in

[2]The choice of $T_1 = 50°C$ and $T_2 = 90°C$ is based on the observation that between these temperatures, the exponential curve is nearly linear. However, a different set of values may be chosen, provided that the temperature rise curve is nearly linear in this interval. But it must be ensured that the temperature limits chosen during calibration and measurement stages are the same.

figure 17 and the inverter is run in open-loop mode. A suitable rms current I_{rms} is established in the load, by adjusting m appropriately. As with the calibration stage, s_{r} is determined.

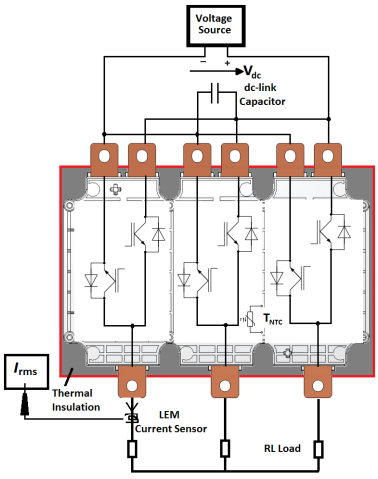
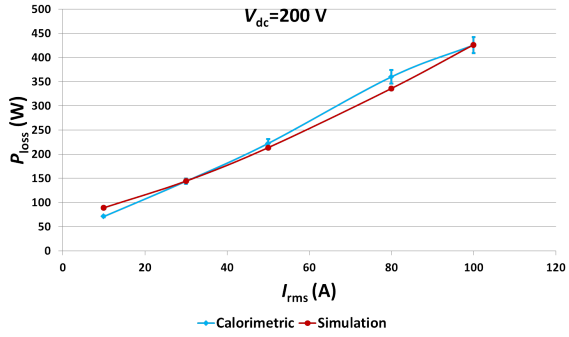
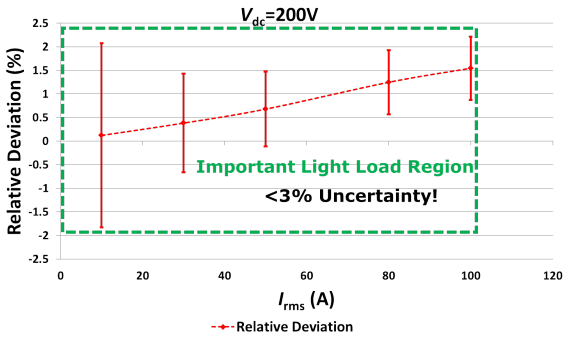


Figure 17: Measurement Stage



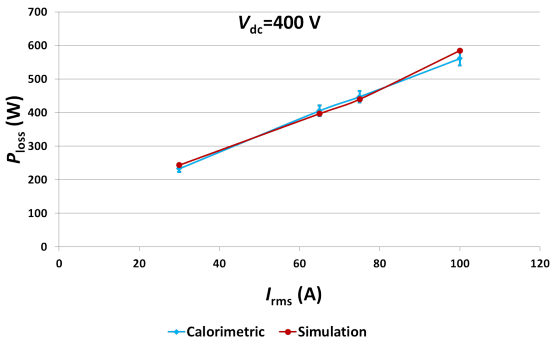
(a) P_{loss}



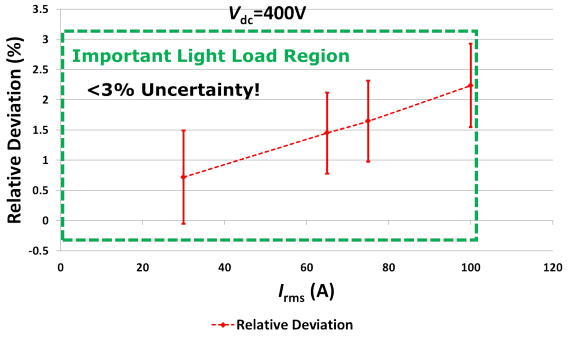
(b) Relative Deviation (%)

Figure 18: Comparison of the calorimetric method with simulations: P_{loss} vs. I_{rms} at V_{dc} = 200V

Now, the inverter losses can be obtained by substituting the measured s_{r} in equation 27 and 28 obtained from the calibration stage. The inverter power losses for different I_{rms} at two different working voltages 200 V and 400 V measured for HPD with the proposed calorimetric method are shown in figures 18(a) and 19(a). These are compared against simulations at the respective points. Also shown are the calculated uncertainties (this topic is described in [3]) for each of the points as bars around the measurement points. The relative deviation between the measurements and the simulations, expressed as percentage, is shown in figures 18(b) and 19(b). Across the entire range of measurement, it can be seen that the simulations are within the tolerance of the calorimetric approach, again validating the behavioral simulation model. Furthermore, in each of these cases, it can be seen that the uncertainty is well below 5%, especially at partial-load which makes the calorimetric method suitable for automotive main inverter applications, particularly for comparing chip technologies whose total power losses may differ in the range of 10-20%.



(a) P_{loss}



(b) Relative Deviation (%)

Figure 19: Comparison of the calorimetric method with simulations: P_{loss} vs. I_{rms} at V_{dc} = 400V

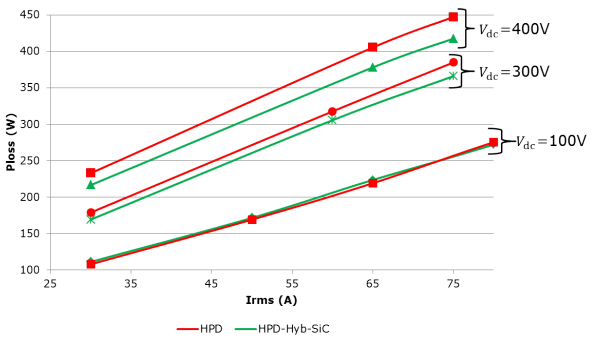


Figure 20: Measured Inverter Losses vs I_{rms} at various V_{dc} for HPD and HPD-Hyb-SiC

7.2.1 Comparison of Measured Inverter Losses for HPD versus HPD-Hyb-SiC

The tests are repeated for different operating points of V_{dc}, I_{rms} and f_{sw} for both the modules. Figure 20 shows a summary of the measured losses as a function of I_{rms} at different dc-link voltages. It can be seen that at V_{dc}=100V, the SiC module has almost the same losses as the Si module, offering no benefit. This is because at low V_{dc}, the switching losses do not contribute much to the total losses, and the conduction losses are dominant. As the conduction losses are higher in the SiC diodes as can be seen from the static curves presented in [2], HPD offers better overall performance than HPD-Hyb-SiC. At V_{dc}=300V, the benefits of SiC become prominent. At V_{dc}=300V, I_{rms}=75 A, there is about 5% reduction in the total losses. This gap widens as we increase V_{dc} as the switching losses become more dominant, and at V_{dc}=400V, 75A, we can see a reduction of around 7%.

8 Conclusions

In this paper, the benefits of replacing the Si diodes of a commercial automotive IGBT module with SiC diodes have been investigated for the main inverter application, maintaining the operating conditions, package and the rest of the system the same, to ensure a fair comparison of the devices without any external influence. A behavioral power loss model, suitable for mission profile analysis, is used to compare the performance of the two modules over several mission profiles. The highest benefit of using SiC diodes is seen for the Artemis Urban drive cycle, where the SiC diodes help reduce the overall losses by 20% at f_{sw}=8kHz. This translates to a saving of around 200Wh of battery energy per 100km. The behavioral model is experimentally verified by comparing it against two independent measurement methods, namely, electrical input output method and a calorimetric method. In each case, the simulation results are found to be within the tolerance of the measurements, thereby validating the simulation model used. At V_{dc} =400V, I_{rms} =75A, the inverter losses were found to be reduced by over 5% with the SiC module, due to the absence of reverse recovery in the unipolar SiC schottky diodes. This reduction is even better at higher dc-link voltages, due to the switching losses becoming more prominent.

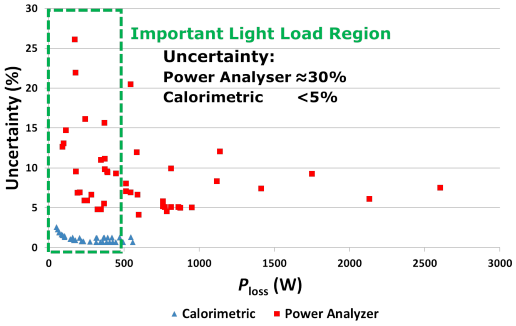


Figure 21: Scatter Plot of the uncertainties (%) at various measurement points for the two methods vs P_{loss}

Further, as summarized in figure 21, it can be concluded that the commonly used power analyser based electrical method has an uncertainty of nearly 30% in the light load condition, mainly due to delays and phase-shifts in the probes. It is to be noted that automotive traction inverters operate most of the time in the light-load condition, which means that the electrical method is not suitable for such applications. The developed calorimetric method outperforms the standard electrical input-output based method and achieves, especially in the important light-load region, a measurement uncertainty of lower than 5%. Furthermore, as the expensive calorimeter is not required for this method, it is nearly as easy to perform as the power-analyser based electrical method. This makes it ideal for comparing device technologies such as Si versus SiC in automotive main inverter applications.

9 Future Work

This work considered the advantages of replacing only the diodes with SiC. However, for higher benefits, it is desirable to replace the IGBTs with SiC MOSFETs, and it would be interesting to investigate the benefits they bring in terms of higher efficiency for different mission profiles. This will considered in a future publication.

References

[1] Ajay Poonjal Pai, Tomas Reiter, and Martin Maerz. Mission profile analysis and calorimetric loss measurement of a sic hybrid module for main inverter application of electric vehicles. In *Integrated Power Packaging (IWIPP), 2017 IEEE International Workshop On*, pages 1–5. IEEE, 2017.

[2] Ajay Poonjal Pai, Tomas Reiter, and Martin Maerz. Characterization and mission profile analysis of a SiC hybrid module for main inverter application of electric vehicles. In *EEHE 2017 Bamberg; Proceedings of*. Haus Der Technik, 2017.

[3] Ajay Poonjal Pai, Tomas Reiter, Oleg Vodyakho, Inpil Yoo, and Martin Maerz. A calorimetric method for measuring power losses in power semiconductor modules. In *EPE 2017 Warsaw; Proceedings of*. ECCE Europe, 2017.

[4] Kimimori Hamada. Great potential of SiC devices for environmentally friendly vehicles. In *APE Automotive Power Electronics*. IEEE, 2015.

[5] Ming Su, Chingchi Chen, Shrivatsal Sharma, and Jun Kikuchi. Performance and cost considerations for SiC–based HEV traction inverter systems. In *Wide Bandgap Power Devices and Applications (WiPDA), 2015 IEEE 3rd Workshop on*, pages 347–350. IEEE, 2015.

[6] Benjamin Wrzecionko, Dominik Bortis, and Johann W Kolar. A 120 C ambient temperature forced air–cooled normally–off SiC JFET automotive inverter system. *IEEE Transactions on Power Electronics*, 29(5):2345–2358, 2014.

[7] SK Singh, F Guédon, PJ Garsed, and RA McMahon. Half-bridge SiC inverter for hybrid electric vehicles: Design, development and testing at higher operating temperature. In *Power Electronics, Machines and Drives (PEMD 2012), 6th IET International Conference on*, pages 1–6. IET, 2012.

[8] Fei Shang, Alejandro Pozo Arribas, and Mahesh Krishnamurthy. A comprehensive evaluation of SiC devices in traction applications. In *Transportation Electrification Conference and Expo (ITEC), 2014 IEEE*, pages 1–5. IEEE, 2014.

[9] Justin K Reed, James McFarland, Jagadeesh Tangudu, Emmanuel Vinot, Rochdi Trigui, Giri Venkataramanan, Shiv Gupta, and Thomas Jahns. Modeling power semiconductor losses in HEV powertrains using Si and SiC devices. In *2010 IEEE Vehicle Power and Propulsion Conference*, pages 1–6. IEEE, 2010.

[10] Hui Zhang, Leon M Tolbert, and Burak Ozpineci. Impact of SiC devices on hybrid electric and plug-in hybrid electric vehicles. *IEEE transactions on industry applications*, 47(2):912–921, 2011.

[11] Burak Ozpineci, Madhu Sudhan Chinthavali, Leon M Tolbert, Avinash S Kashyap, and H Alan Mantooth. A 55-kW three-phase inverter with Si IGBTs and SiC schottky diodes. *IEEE Transactions on industry applications*, 45(1):278–285, 2009.

[12] M Chinthavali, Leon M Tolbert, Hui Zhang, Jung H Han, F Barlow, and Burak Ozpineci. High power SiC modules for HEVs and PHEVs. In *Power Electronics Conference (IPEC), 2010 International*, pages 1842–1848. IEEE, 2010.

[13] Timothy Junghee Han, Jim Nagashima, Sung Joon Kim, Srikanth Kulkarni, and Fred Barlow. Implementation of a fully integrated 50 kW inverter using a SiC JFET based six–pack power module. In *2011 IEEE Energy Conversion Congress and Exposition*, pages 3144–3150. IEEE, 2011.

[14] Florian Hilpert, Klas Brinkfeldt, and Stefan Arenz. Modular integration of a 1200 V SiC inverter in a commercial vehicle wheel–hub drivetrain. In *Electric Drives Production Conference (EDPC), 2014 4th International*, pages 1–8. IEEE, 2014.

[15] Product brief HybridPACK drive. *Infineon AG*, 2014.

[16] Datasheet 5th generation thinQ 650V SiC schottky diode IDW40G65C5, author=, journal=Infineon AG, year=2013.

[17] Ajay Poonjal Pai, Tomas Reiter, and Martin Maerz. A new behavioral model for accurate loss calculations in power semiconductors. In *PCIM Europe 2016; Proceedings of*. VDE, 2016.

[18] Ajay Poonjal Pai, Tomas Reiter, and Martin Maerz. An improved behavioral model for loss calculations in automotive inverters. In *EEHE 2016 Wiesloch; Proceedings of*, pages 412–427. Haus Der Technik, 2016.

[19] Lem current transducer IT 200-S ULTRASTAB, author=, journal=LEM, year=2014.

[20] Lem current transducer LF 510-S, author=, journal=LEM, year=2015.

[21] Yokogawa WT1800 getting started guide, author=, journal=Yokogawa, year=2015.

[22] C. Xiao, G. Chen, and W. G. H. Odendaal. Overview of power loss measurement techniques in power electronics systems. *IEEE Transactions on Industry Applications*, 43(3):657–664, May 2007.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 342-347 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Signal-Based Metamodels for Predictive Reliability Analysis and Virtual Testing

Veit Bayer[*,1], Stephanie Kunath[1], Roland Niemeier[1], Jürgen Horwege[2]

[1]*Dynardo GmbH, Weimar, Germany*

[2]*Airbus Operations GmbH, Bremen, Germany*

ARTICLEINFO

Article history:
Received: 29 November, 2017
Accepted: 19 January, 2018
Online: 02 February, 2018

Keywords:
Metamodel
Surrogate model
Dynamics
Random field
Aircraft
High lift
Digital twin
Virtual test

ABSTRACT

In the current industrial development, an increasing number of sensors is applied for monitoring of any kind of appliances and machines. A predictive mathematical model allows for realistic assessment of the health state of the appliance, indication of service requirements, as well as control of the appliance, serving as so-called digital twin on a control device.

Completely analogous modeling can be used for virtual testing, i.e. a (partial) substitution of physical experiments. The software-in-the-loop model provides realistic feedback to the physical specimen on the test rig and helps to increase the representativeness of the experiment and to reduce costs.

In a joint research project partially sponsored by the German Federal Ministry of Economics and Energy, the authors developed an approach for meta-modeling of dynamic systems. While the modeling process is fed by results from sophisticated simulations, or even test results as input data, the resulting model can be used for fast stochastic analyses as well as software-in-the-loop in dynamic real-time experiments. The approach was verified on tests of an aircraft high lift system.

1 Motivation and Objective

The prediction of test results is especially important for very expensive tests like in aerospace industrial applications, e.g. testing of high lift systems. As a matter of costs and manageability, the test specimen usually represents only a partial system. A realistic feedback of reactions from the remaining system by software driven actuators increases the representativeness of test results. Sophisticated procedures for realistic modeling of physical systems, e.g. algorithms for multi-body dynamics simulation, are available, but they require significant computation time and lack the ability of real-time performance. Therefore, there is a need to reduce computational cost maintaining high fidelity modeling.

Purely mathematical models can be established which are fast to compute, yet maintain accurate results. The authors' approach, which has been presented initially in [1], yields meta-models that approximate the dynamic response of the tested object. Models are built on the basis of physical simulations that represent the test results accurately, but require a computation time that would be prohibitive as for the applications indicated above. The results of real experiments can be used as input data as well. The proposed meta-models, having short response times, enable fast robustness analysis to assess the influences of uncertain parameters such as damping or friction and can be used as software-in-the-loop even in real-time dynamic experiments [2, 3], where often fast but simplified physics-based models are applied [4]. The requirements for development of the methodology, which will be presented in the following sections, are summarized as: speed in performance, flexibility to change of parameters, versatility and accuracy.

The combination of physical testing and virtual testing serves several purposes: evaluation of the test rig without risk of damage; substitution of parts of the tested system —or augmentation of the existing test—; more realistic and more complex tests and reduction of costs.

[*]Veit Bayer; Dynardo, Steubenstr. 25, D-99423 Weimar, Germany; Ph. +493643900847; E-mail veit.bayer@dynardo.de

https://dx.doi.org/10.25046/aj030141

Other applications of this approach can be thought of: fast meta-models can be implemented as digital twin on data acquisition and control devices for online monitoring of a system; in the development of a product, fast but accurate models are important for stochastic analyses such as predictive reliability assessment.

2 Methodology

Basis of the methodology proposed here is Dynardo's algorithm *Metamodel of Optimal Prognosis* (MOP). The procedure fits the best available model to given data points and avoids so-called over-fitting. It first starts a filtering of parameters by statistical criteria. Input parameters with negligible influence on the observed results are canceled from the data set, thus the dimension of the problem can be reduced effectively. Several model approaches, such as polynomials, moving least squares, kriging, are built up and tested by cross validation. The resulting *Coefficient of Prognosis* (CoP) is the complimentary value to the sum of squared residuals over the variation of result data. The CoP gives information, what amount of data variation is explainable by the meta-model, based on independent test data. The MOP result is the chosen model out of the set of available models with the largest CoP. The MOP is also used for sensitivity analysis [5, 6]. Conditional variances are computed by holding systematically one parameter at fixed values, indicating the relative contribution of this parameter to the total variation of the response.

The MOP deals with scalar response quantities only. For the intended application, results are time series, e.g. from a multi-body dynamics simulation. Hence a representation of the time series by scalar values has to be found, while avoiding to adopt the total set of discrete time steps. The approach makes use of a specific topic in probability theory, namely the random fields methodology [7, 8]. A random field is a quantity defined on a spatial domain, where the value at any point of observation is a random variable. Here, the domain is time instead of space. The training data for the model are produced by first sampling input values, either by design of experiments or quasi-random sampling, then computing the results for each input data set. The time-dependent results are then interpreted as random process or 1-D random field.

The key to a parametrization of the random process is the eigenvalue decomposition of the covariance matrix $\mathbf{C}_{XX}$ of the discretized time series $\mathbf{X}$,

$$\mathbf{\Psi}^T \mathbf{C}_{XX} \mathbf{\Psi} = \mathrm{diag}\{\lambda_i\}, \tag{1}$$

wherein $\mathbf{\Psi}$ is the matrix of eigenvectors, and $\mathrm{diag}\{\lambda_i\}$ holds the eigenvalues of the covariance matrix. From this, the so-called *spectral representation* of the random field can be derived [9]. Assuming that X are normal distributed with zero mean values (which can be subtracted for the analysis and added later again for synthesis of time series), new random variables Y are defined as

$$\mathbf{Y} = \mathbf{\Psi}^T \mathbf{X}, \tag{2}$$

which are normal distributed, independent, zero-mean and with a standard deviation given by

$$\sigma_{Y_i} = \sqrt{\lambda_i}. \tag{3}$$

For synthesis of the original time series $\mathbf{X}$, one makes use of the Karhunen–Loève series expansion

$$\mathbf{X} = \mathbf{\Psi} \mathbf{Y}. \tag{4}$$

A typical property of eigenvalue solvers is, that eigenvalues are stored in descending order. Since the eigenvalue of order i defines the amount of variation contributed by parameter Y_i to the total variation of the data, this gives a criterion for truncation of the series expansion and therefore a drastic reduction of the dimension [10, 11].

Summarizing, for generation of a time series $\mathbf{X}$ one needs the modal base $\mathbf{\Psi}$ of the covariance matrix, which we may call the set of "shape functions" in the following, and the respective "amplitudes" Y_i. Figure 1 shall illustrate the series expansion of (4) with an added mean value signal.

The shape functions are an unchangeable property of the data. Properties of the parameters Y_i have to be determined such that, the series expansion optimally represents the physical time series $\mathbf{X}$. For this purpose, MOP is applied to a training data set, from which the corresponding sample of amplitudes is calculated. This leads to the *Field-Metamodel of Optimal Prognosis* (F-MOP). Figure 2 shows the flow of analysis. Using these dynamic meta-models, we obtain a simplified and reduced parametric of the dynamic signal based on a statistical meta-model. The user does not need to find a parametrization himself.

In analogy to the sensitivity analysis of scalar data [5, 6], the Coefficient of Prognosis of the F-MOP can be plotted against the time axis, called F-CoP here. Moreover, sensitivity measures over time can be computed, such that it is possible to assess the model quality and the relative influence of input parameters on the response, locally within the entire observed time range. The procedure is realized by connecting the functionalities of the programs offered by Dynardo, optiSLang and Statistics on Structures (SoS).

3 Application Example: Aircraft High Lift System

The methodology which is described in the previous section is validated by the example of an aircraft high lift system. Instead of real experiments, virtual tests using a detailed model serve for generating the training data. The simulation model in MSC Adams/Flex comprises the inboard flap, outboard flap, transmission and actuators as well as the test rig. With that, model wing position, flap positions, loadings as well as

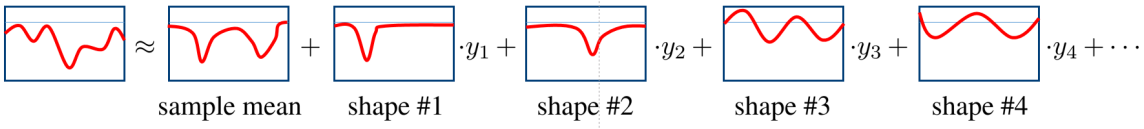


Figure 1: Illustration of Karhunen–Loève series expansion of a signal.

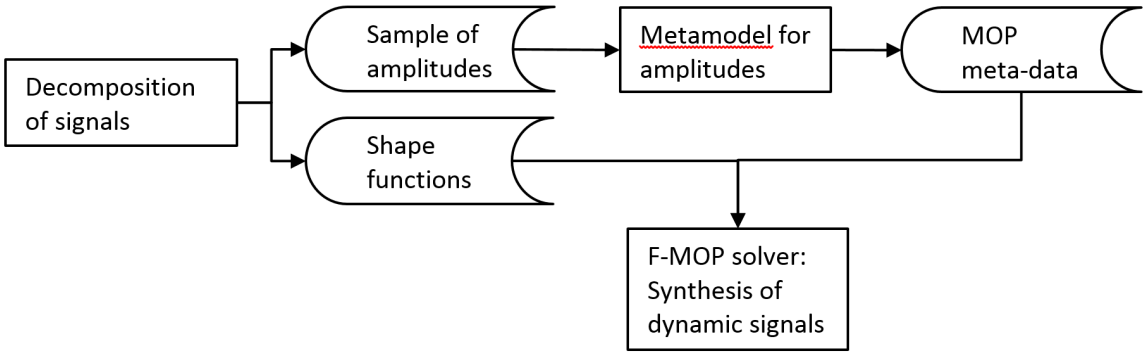


Figure 2: Overview of the approach of decomposing signals and generating meta-models for dynamic signals.

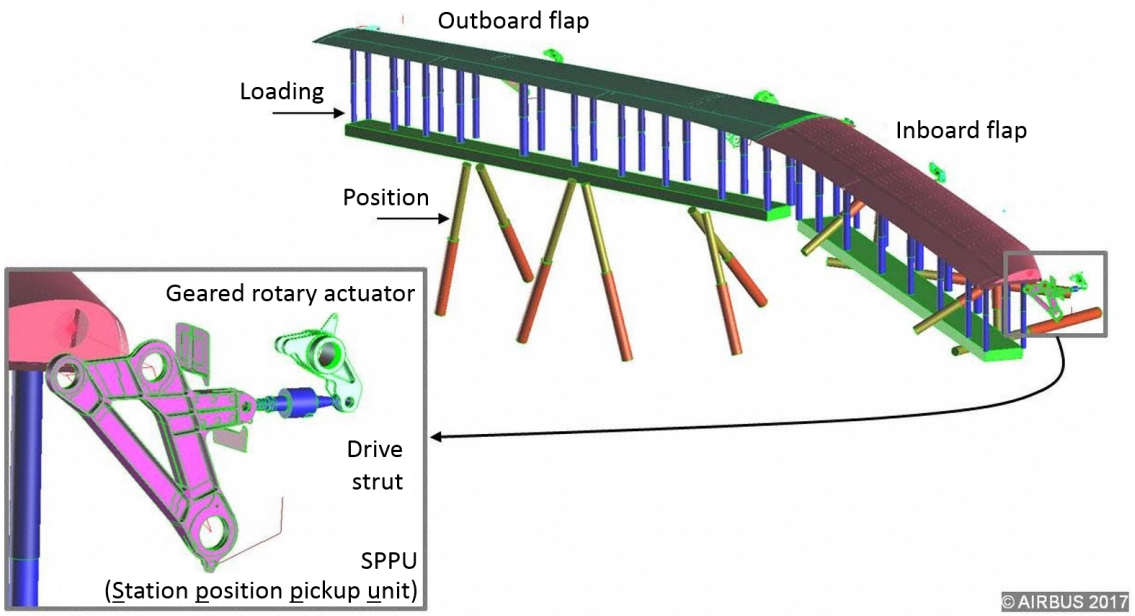


Figure 3: Adams/Flex model of the high lift system on a test rig.

actuator action and backlash can be simulated. Since friction is taken into account in the simulation, the model is actually non-linear. Figure 3 visualizes the tested system.

For the virtual test, wing and flap positions and loadings are given. The test case is a simulated rupture of two actuators simultaneously, one at the inboard flap, one outboard. As uncertain parameters, the stiffness of all actuators, backlash of all actuators, damping and friction parameters are sampled as input to the meta-modeling procedure. 200 parameter sets are simulated using MSC Adams/Flex, yielding signals over time of the dynamic responses. The responses observed are all actuator moments, drive strut forces and angles at the station position pickup units (SPPU). For these responses, dynamic meta-models shall be built.

4 Discussion of Results

Figure 4 –top– shows 4 typical signals picked out of the whole sample, here of the moment at actuator #2. Below, the prediction capability expressed as F-CoP is plotted. It can be observed that, the Coefficient of Prognosis of the model F-CoP [Total] is able to reach nearly 100% at the peaks of the signal, while its value drops where the oscillation passes the stationary value. At such points, scatter in data is rather numerical noise than information and cannot be explained by the meta-model. The relative influences of input parameters can be observed, too (largest: F-CoP[DV_Fric_Mue_PP_dyn] for Coulomb friction parameter).

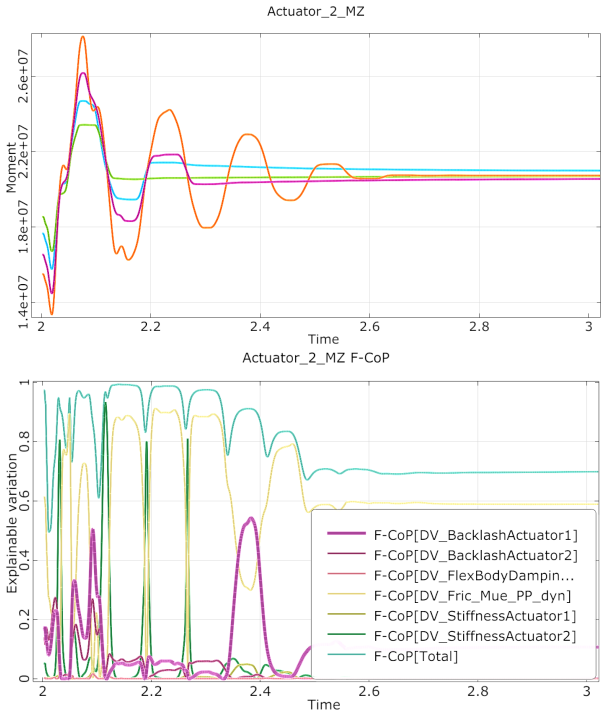


Figure 4: Top—Sample time series of moments at actuator #2. Bottom—Prediction capability of the model for the time dependent signal (F-CoP [Total]) and the impact of each input parameter on the total variation of the signal.

Another signal, the angle at SPPU #2, is examined in the same way, results are displayed in Figure 5. There is a constant shift for each sample. The F-CoP[Total] values are near 100% throughout the entire time range. The model quality is considered excellent by means of the statistical F-CoP criterion.

Finally, the original simulated time series and the series synthesized with help of the field meta-model shall be compared. Again, a few time series are arbitrarily picked from the whole sample. In Figure 6, the signals of the moment at actuator #2 are compared, and Figure 7 shows the comparison for the angles at the station position pickup unit #2. The left plots in both figures are the original simulation results obtained by MSC Adams/Flex, which serve as reference. The right plots are the signals which were synthesized by the field meta-models using the same input parameters as for the original simulations. The coincidence of original signals and meta-model results is excellent, particularly in the peaks. When the signals tend to be damped out, some artificial oscillations in the field meta-model results can be observed. These will vanish if the series expansion (refeq:KLseries) is truncated at a later position, thus more shape functions will be taken into account.

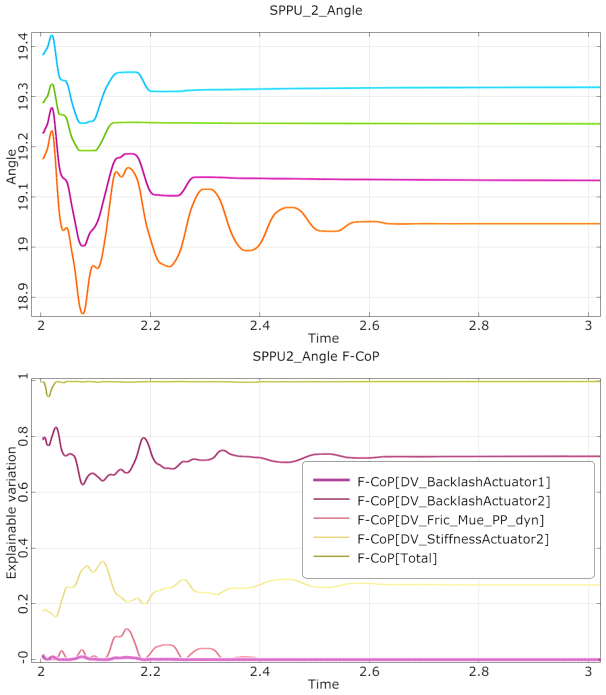


Figure 5: Top—Sample time series of angles at SPPU #2. Bottom—Prediction capability of the model for the time dependent signal (F-CoP [Total]) and the impact of each input parameter on the total variation of the signal.

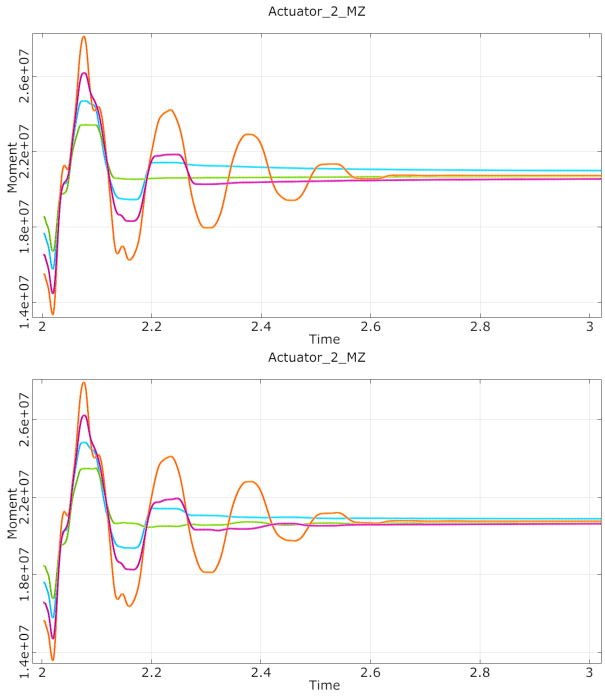


Figure 6: Sample time series of moments at actuator #2. Top—original results obtained from Adams/Flex. Bottom—synthesized signals by F-MOP.

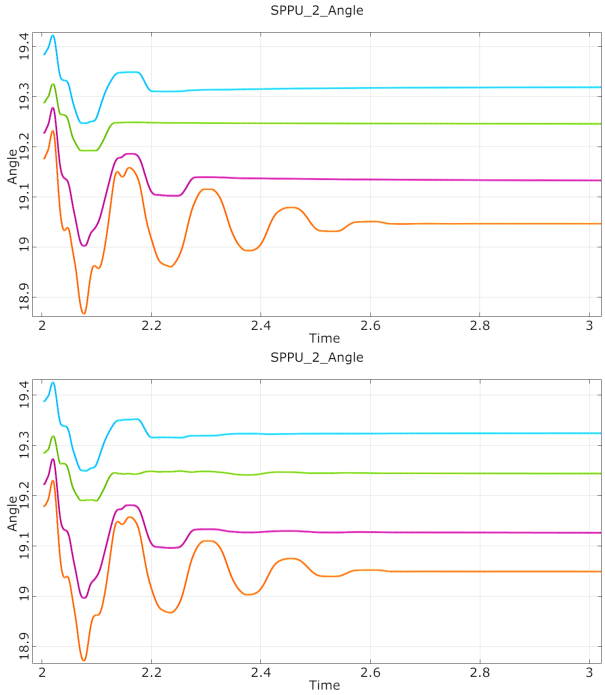


Figure 7: Sample time series of angles at SPPU #2. Top—original results obtained from Adams/Flex. Bottom—synthesized signals by F-MOP.

The proposed procedure has been applied also to other applications, e.g. for the parameter identification of non-linear material models in finite element analysis. The "signal" here is the load–displacement curve of a test specimen. In previous approaches [12], the signal was discretized into few equidistant steps, then meta-models were established by MOP for the single values on the curve. The squared residuals at these discrete points were used as criteria in a fast optimization procedure using MOP. For longer signals this approach is infeasible, moreover, it requires manual interference to identify characteristic points or relevant ranges. With the proposed approach, this is not necessary anymore. The field meta-model is valid for all points according to the raster of the original data, and one easily sees in which regions the parameters have what amount of influence.

5 Summary and Outlook

A new procedure is proposed in this article for meta-modeling of dynamic (time series) signals. It is based on the decomposition of a sample set of signals into a series consisting of shape functions scaled by amplitudes for each summand. The algorithm Metamodel of Optimal Prognosis is then applied to find the best fitting model for each amplitude. The Coefficient of Prognosis, which does an assessment of models with the statistical method of cross validation, is used as the selection criterion. It is possible to assess the model prediction capability even locally along the time axis. Moreover, sensitivity measures reveal the relative influence of parameters (as input to the dynamic analysis or experiment), also locally along the time axis.

The procedure has been applied to an aircraft high lift system. The training data were generated by virtual experiments, i.e. multi-body dynamics simulations. The surrogate models proved to be very accurate, by the CoP criterion as well as by direct comparison of the reference to the model time series. Unlike the high-fidelity simulation, the meta-models are very fast to compute, allowing e.g. for fast stochastic analyses in the scope of predictive reliability analysis. The models were also successfully implemented as software-in-the-loop into an experimental environment for real-time dynamic tests at the Airbus site.

Further developments are planned which will focus on more detailed model assessment and model improvement, locally in the time and parameter space. The range of application for this procedure is not limited to the above mentioned. Signals can be any xy-data, so fast surrogate models can be obtained, e.g., also for spectral analyses both in the low and high frequency domains, stress-strain curves of a material law etc. The field meta-models can also serve as digital twins which, fed by sensor data, monitor or control electronic appliances or machines.

Acknowledgments This work was partially financed by the German Federal Ministry of Economics and Energy, call LUFO 5.1, grant number 20Y1301E, which is gratefully acknowledged by the authors.

The authors also express their gratitude for the contributions by Thomas Töpsch, Tobias Ulmer, Achim Lenz, Airbus Operations, Bremen; Sönke Klostermann, Mario Cappitelli, Airbus Group Innovations, Hamburg; Viktor Lebsak, MSC Software, München and Michael Neumann, P3 group, Hamburg.

References

[1] S. Kunath, V. Bayer, R. Niemeier, "Predictive reliability with signal based meta-models" in 18^{th} International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE), Dresden Germany, 2017.

[2] A. Forrester, A. Sóbester, A. Keane, Engineering Design via Surrogate Modelling, Wiley 2008.

[3] D. Vogt, "Effektive Visualisierung Stochastischer Simulationen", Der Andere Verlag 2010.

[4] V. Bayer, U.E. Dorka, U. Fllekrug, J. Gschwilm, "On real-time pseudo-dynamic sub-structure testing: algorithm, numerical and experimental results" Aerospace Science and Technology, **9**(3), 223–232, 2005.

[5] T. Most, J. Will, "Metamodel of Optimal Prognosis – an automatic approach for variable reduction and optimal metamodel selection" in Proc. Weimar Optimization and Stochastic Days 5.0, Weimar Germany 2008.

[6] T. Most, J. Will, "Sensitivity analysis using the Metamodel of Optimal Prognosis" in Proc. Weimar Optimization and Stochastic Days 8.0, Weimar Germany, 2011.

[7] E. Vanmarcke, Random Fields: Analysis and Synthesis, MIT Press 1983.

[8] C. Bucher, Computational Analysis of Randomness in Structural Mechanics, Structures and Infrastructures Book Series Vol. 3, CRC Press 2009.

[9] R. Ghanem, P.D. Spanos, Stochastic Finite Elements – a Spectral Approach, Springer 1991.

[10] C.E. Brenner, "Ein Beitrag zur Zuverlässigkeitsanalyse von Strukturen unter Berücksichtigung von Systemuntersuchungen mit Hilfe der Methode der Stochastischen Finite Elemente", Ph.D. (Dr. tech.) Thesis, Leopold Franzens Universität Innsbruck, 1995.

[11] V. Bayer, J. Will, "Random Fields in Robustness and Reliability Assessment of Structural Parts" in 15. VDI Kongress Berechnung und Simulation im Fahrzeugbau SIMVEC (in German with English abstract), Baden-Baden Germany 2010.

[12] T. Most, "Effiziente Parameteridentifikation fr numerische Simulationsmodelle", in NAFEMS Konferenz: Berechnung und Simulation, Bamberg Germany, 2014 (in German)

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 348-351 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

LabVIEW-based data acquisition system for Diode I-V Characterization

Nor Shaida Mohd Saufi*, Nurul Syafiqah Yap Abdullah, Mohd Ikhwan Hadi Yaacob

Faculty of Science Mathematics, Universiti Pendidikan Sultan Idris, 35900, Malaysia

ARTICLE INFO

Article history:
Received: 15 November, 2017
Accepted: 07 January, 2018
Online: 02 February, 2018

Keywords:
LabVIEW
Data acquisition system
I-V Characteristics
Diode

ABSTRACT

This work describes computerized data acquisition system for current-voltage (I-V) characteristics of diodes. The proposed system consist of a personal computer pre-installed with LabVIEW, two units of power supply, an interface board, current and voltage sensors. Two devices under test (DUTs) are selected namely 1N4007 silicon and 1N34 germanium diodes. The current flow through the circuit and voltage across the diode is measured and I-V characteristic is successfully plotted. When data acquisition interrupted by the user, real time I-V curve with least-square fitting is plotted. The LabVIEW was employed primarily to display data from sensors and storing collected data for further post-measurement processing. Estimated forward voltage drop from plotted I-V curve was then compared with the published value from manufacturer's datasheet. Comparison revealed that developed system has been successfully characterized both DUTs base on the forward voltage drop obtained from the plotted I-V curve. Furthermore, proposed system architecture offers extra flexibility where stored data can be manipulated with minimum programming efforts. The application of the proposed system can also be extended for in-situ device characterization in various circuits.

1. Introduction

Rapid advancement in computerized measurement system around the world has seen the introduction of fast, simple, user-friendly and reliable solutions found in various applications. In recent years, rapid progress of the virtual instrument platform such as LabVIEW graphical programming language has simplified programming effort and produce simple yet highly-productive real-time computerized measurement system.

This paper is an extension of work originally presented in 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE) [1]. In this report, we discussed in detail a system to determine I-V characteristics curve of germanium diode.

Proposed computerized I-V characterization system employed LabVIEW as the acquisition tool to obtain the data, analyze and display the result. Ease of use is one of our priority in this project. However, the complete functionality of LabVIEW is a plus factor that also satisfied the requirement of this project [2]. Virtual Instruments (VIs) in LabVIEW is utilized to construct a Graphical User Interface (GUI) as the time saving is the fair trade-off for the full low level control of other conventional programming languages such as C/C++, Visual Basic, and Matlab. Its graphical nature makes it ideal for measurement and data acquisition [3,4].

Simplifying data acquisition process and decreasing test time without scarifying the accuracy are primary goals in this project. Semiconductor diode is used in all sorts of electrical and electronic system. The electrical characterization of diode has been an important subject for more than half a century [5]. Semiconductor diode is an important device that exhibit non-linear relationship of I-V characteristics. In order to visualize the relationship, it is common to plot current versus voltage on graph. However, it is much more convenient to display the complete curve in an automated computer-based acquisition system.

2. System Description

This paper describes a LabVEW-based measurement method to plot I-V characteristics of semiconductor diodes. Figure 1 shows a block diagram of the main components of the measurement system that comprised mainly a personal computer with LabVIEW software, voltage and current sensors, interface board to be used between the sensors and computer, power supplies, and electronic circuit. In this research, LabVIEW 2012 was used in Windows 7 operating system. More precise sensors were used to measure voltage and current. The current sensor used

*Nor Shaida Mohd Saufi, Email: norshaidasaufi@gmail.com

was the Phidget CE-IZ02-32MS2-0.5 DC Current Sensor 0-1A powered by an external 12V power supply. The voltage sensor is Phidget Precision voltage sensor 1135 which measure differential voltage between input terminal and its output terminal. The current and voltage sensor was interfaced to personal computer through Phidget interface board 8/8/8 through Universal Serial Bus (USB) cable. Basically, the main components of the system and circuit used in this project is similar as described in [1] but we have used more sophisticated and better performance voltage and current sensors in terms of its measuring range and accuracy.

3. LabVIEW software

3.1. Block diagram

There are two windows for user to works on which are block diagram and front panel. LabVIEW is a programming environment which in block diagram, blocks representing functions, icons representing variables and lines representing data flows pass between different functional nodes [4]. LabVIEW is compatible with most of other hardware such as Phidget devices therefore they have their own blocks representing function. The communication between Phidget devices and LabVIEW software has been simplified using robust Application Program Interface (API) library which is accessible online for free. Figure 1 shows the block diagram of formula node in LabVIEW that was used to convert analog sensor value of voltage and current sensors to its corresponding value in volts and miliAmperes.

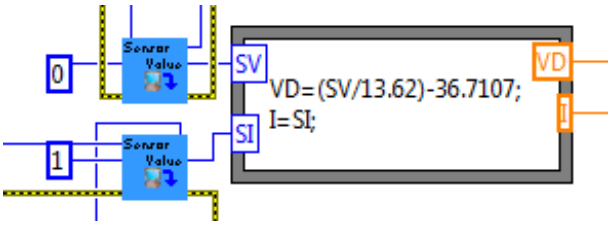


Figure 1. Formula node to write formula used to convert analog sensor value to voltage and current values

Particularly, Lev-Mar least square fitting method is expounded in the system which fit best the non-linear relation of diode. Figure 2(a) shows block diagram in LabVIEW for diode when user switch Curve fitting slide to Non-Linear and in Figure 2(b) is the block diagram when user switch slide to Linear. The saved data file in .csv can be accessed and plotted in Compare Graph when user clicked 'file path' and choose file location. The block diagram of the flow pass is shown in Figure 3.

Figure 3 shows block diagram that its function to read data from spreadsheet or .csv data file to be plot and compare in graph. The functional block can read and display multiple plot of germanium and silicon diode.

3.2. Graphical User Interface (GUI)

Front panel window is used as Graphical User Interface (GUI) of the system for user to interact with the program. Controls such as push button START, STOP, SAVE, PLOT and etc. as shown in Figure 4 are for user action to make the data displayed in indicator fields such as voltage, current value and also graphs. Before start the acquisition, user must make sure to check what is the electronic device to be tested. As we performed the measurement for I-V characteristics of diode, the Curve Fitting slide must be set to Non Linear. Next, when user clicked RUN button to start performing the acquisition, the LED of RATIOMETRIC STATE button is appeared green when the interface board is properly connected to computer and appeared red when it is not connected.

The voltage and current value appeared automatically and plotted on the I-V Characteristic graph. The fitting of I-V plot is plotted in real time until user push the STOP button to abort the execution. The graph provided a cursor for user to hover over the plotted line to find forward voltage drop of the diode as shown in Figure 4. The GUI gave user option in Curve fitting slide button to choose between linear and non-linear regression depends on what the characteristics of device the user testing.

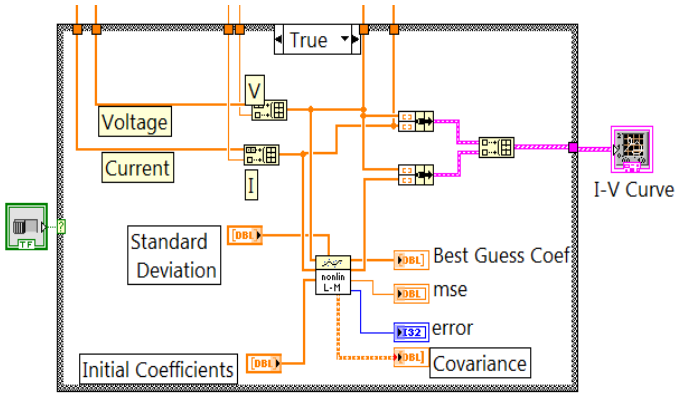


Figure 2(a). Block diagram of linear curve fitting

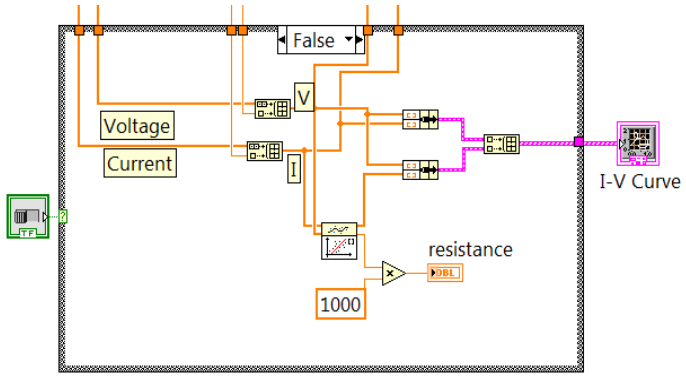


Figure 2(b). Block diagram of non-linear curve fitting

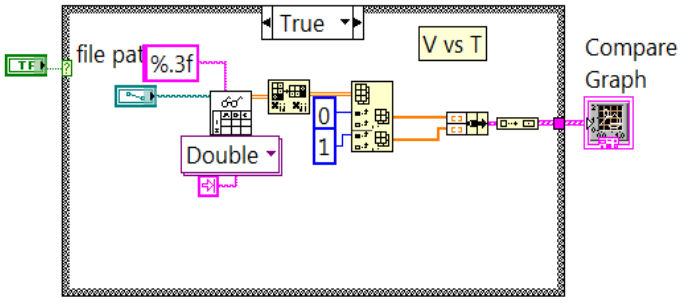


Figure 3. Functional block to read and plot graph from data file

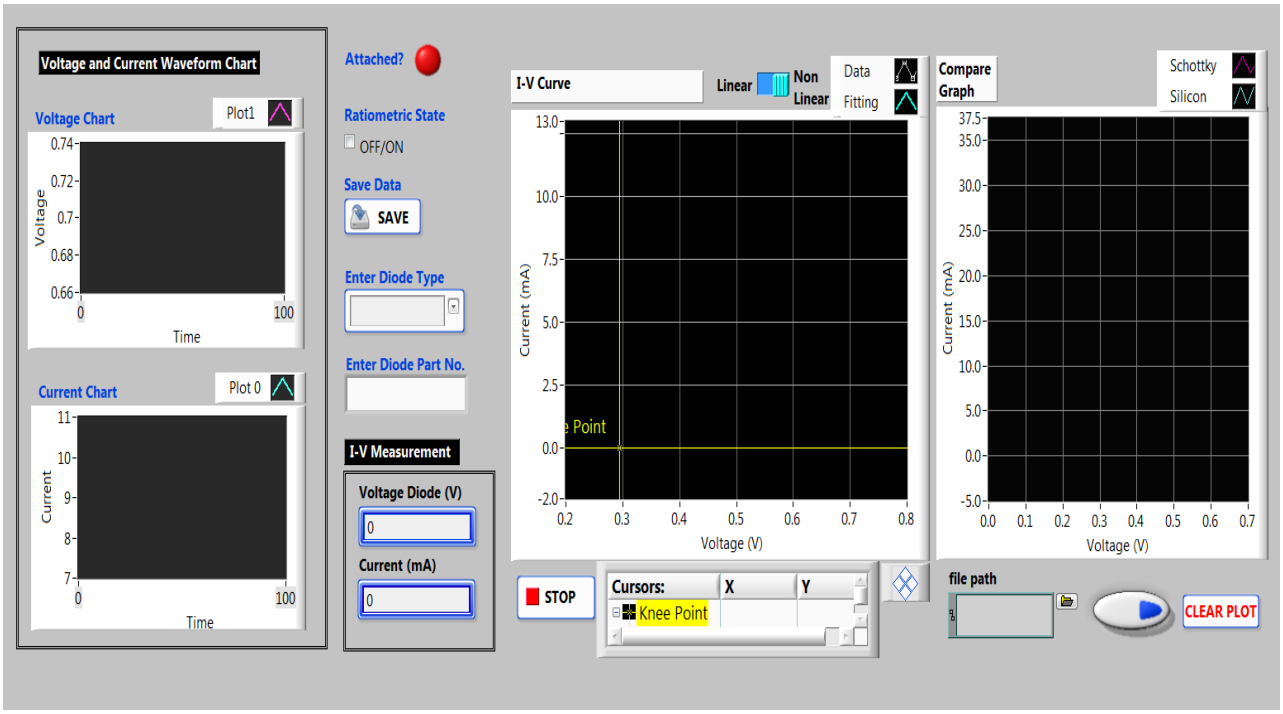


Figure 4. Graphical User Interface (GUI) developed in LabVIEW software

4. Analysis & Discussion

Recently in [1], silicon and Schottky diode has been studied which the system has shown the non-linear relationship of I-V characteristics. Upon studying the I-V graph, forward voltage drop of both diodes has been clarified which 0.7 volts for silicon and 0.4 volts for Schottky diode.

In this paper, the system has been tested on same silicon diode and characteristic of germanium diode is explored. The I-V characteristics graph plotted in Figure 5 can be seen that after the forward voltage drop, the current increases very rapidly where the device starts to conduct for both DUT. Below this voltage, current is less than 1% of maximum rated value of diode current [6]. Rated peak forward current for 1N34 and 1N4007 diode for most brand are 150mA and 30A [6,7]. Voltage increase slightly due to small internal dynamic resistance as the current change after the forward voltage drop is larger than the voltage change. The 1N34 germanium diode has much less defined result. The forward voltage drop is 0.4V which it is still shown low forward voltage drop. This germanium type of diode have low forward voltage drop which is fast switching speed and found uses in television and crystal radio application.

The silicon diode always expected to have value of forward voltage drop which is approximately 0.7 volts. The forward voltage drop was measured as 0.7 volts which this has support the theory behind its design. A germanium diode such as 1N34 have typically 0.3 volts forward voltage drop or low forward voltage drop means they are much more efficient compared to silicon diode. However, silicon diodes are resistant to heat, and better in terms of processing and stability. The results of this experiment appear to satisfy the theory behind diodes.

5. Conclusion

In this paper, a system for determination of current-voltage characteristics of diode has been developed and implemented. Based on LabVIEW software platform, this system measure current flow through the circuit and voltage across the diode to plot I-V characteristics of diode tested. From I-V plotted, forward voltage drop of silicon and germanium diode has been successfully determined. The possible research direction may include determination of voltage-current characteristics of any other two terminal components such as resistor or transistor. The ideality factor of diode can also be estimated by taking the natural log of current versus voltage plot.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors want to thank to the Ministry of Higher Education (MoHE), Malaysia, through the RAGS/1/2015/STO/UPSI/03/1 Research Acculturation Grant Scheme (RAGS), Universiti Pendidikan Sultan Idris (UPSI) for the financial support provided to this work.

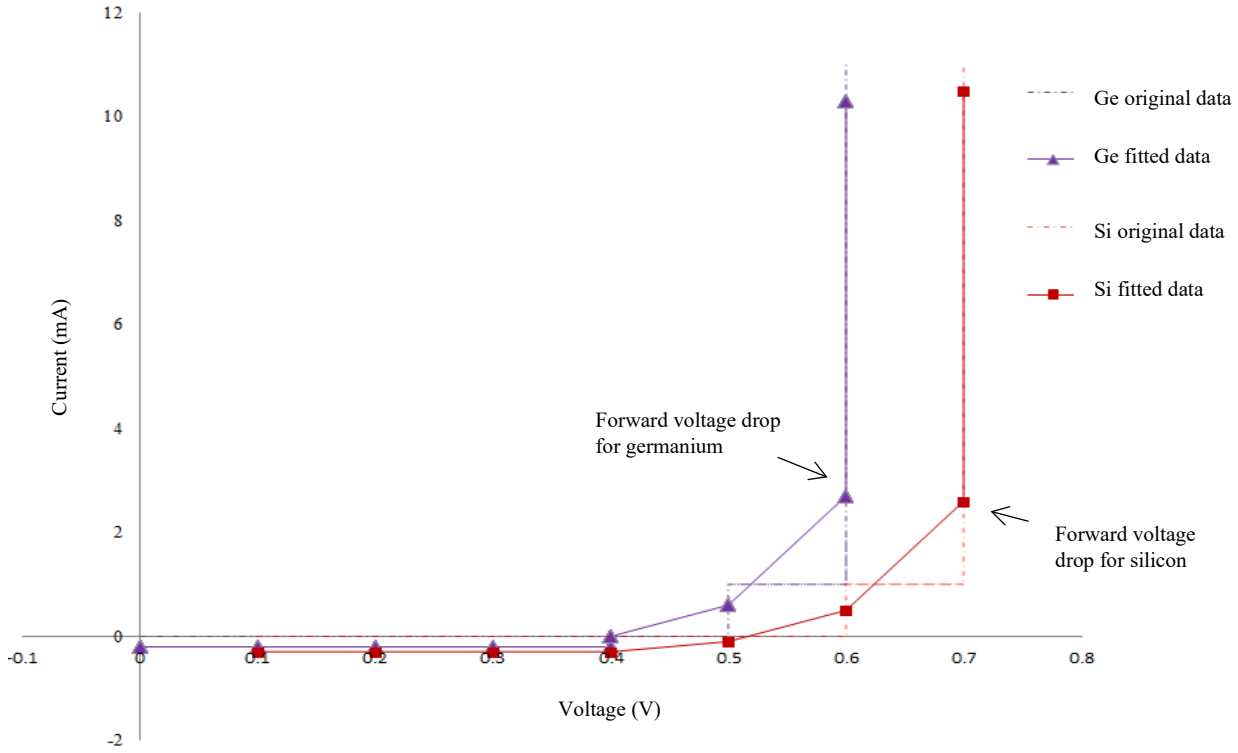


Figure 5. I-V characteristics graph of silicon and germanium diode

References

[1] N.s., M. S., Abdullah, N., & Yaacob, M. (2016). "Automated measurement system for diode I–V characterization," in 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE),pp. 498-501,2016.

[2] Sumathi, S., Surekha, P., & Surekha, P. (2007). LabVIEW based advanced instrumentation systems (Vol. 728). Berlin: Springer.

[3] J. Travis, J. King, LabVIEW for Everyone: Graphical Programming Made Easy and Fun, 2006.

[4] Bitter, R., Mohiuddin, T., & Nawrocki, M. (2006). LabVIEW: Advanced programming techniques. Crc Press.

[5] W. Shockley, "The theory of p-n junctions in semiconductors and p-n junction transistors," Bell Syst. Tech. J., vol. 28, p. 435, 1949.

[6] Bakshi, A. G. U. (2009). Basic Electronics. Technical Publication

[7] Semtech, "Point Contact Germanium Diode," 1N34A datasheet, n.d. {Retrieved from http://pdf1.alldatasheet.com/datasheet-pdf/view/42335/SEMTECH/1N34.html

[8] Vishay, "General Purpose Plastic Rectifier," 1N4001 thru 1N4007 datasheet, 2017 [Revised Feb. 2017].

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 352-356 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Innovative design with learning reflexiveness for developing the Hamiltonian circuit learning games

Meng-Chien Yang *, Hsuan-Yu, Chiang

[1] *Department of Computer Sciences and Communication Engineering, Providence University, Taichung, 43301, Taiwan ROC*

ARTICLE INFO

Article history:
Received: 30 November, 2017
Accepted: 07 January, 2018
Online: 02 February, 2018

Keywords :
Game-based learning,
Graph theories,
Learning reflexiveness,
Hamiltonian circuit

ABSTRACT

In this study, we use a new proposed framework to develop the Hamiltonian circuit learning games for college students. The framework is for enhancing learners' activities with learning reflexiveness. The design of these games is based on this framework to achieve the targeted learning outcomes. In recent years, the game-based learning is a very popular research topic. The Hamiltonian circuit is an important concepts for learning many computer science and electric engineering topics, such as IC design routing algorithm. The developed games use guiding rules to enable students to learn the Hamiltonian circuit in complicate graph problem. After the game, the learners are given a reviewing test which using the animation film for explaining the knowledge. This design concept is different from the previous studies. Through this new design, the outcome gets the better learning results under the effect of reflection. The students will have a deeper impression on the subject, and through self-learning and active thinking, in the game will have a deeper experience.

1. Introduction

The study of the paper is the research of the game-based learning technology. The game-based learning is an emerging research topics of the learning technology. Developing a video game as a learning tool might include many different tasks and considerations. Basically, how the game can be effective for learning is the main concern problem. This problem can arouse many challenging research problems such as the learning model of the game, the foci of the game, the players' attitude of the game and the game development. Many studies focused on the game developments such as the development of a new gaming technology, the usefulness of the gaming tool proposed by the game company and the development process of the learning game. Other studies might focus on the effectiveness of the game such as the learning process, the learners' attitudes. Moreover, some complicate topics of the game industry such as the immersive experience of the players brought research interests. Traditionally, the game for learning the specific subject is regarded as the serious game. Many different serious games have developed for different purpose of training and learning. Although the serious games have been used as the tool for education, most studies of the serious games remained on development of the functions of the game. Then how these games are effective and helpful for learning? This could be another challenging research topic. Our research goal is to find how to develop the effective learning method for game-based learning.

In this paper, a new framework for developing the learning game is proposed. Using advanced information technology, the design of digital games is booming in many applicable developments. Because of the unique interactive interesting and attractive features in digital games, it can arouse the students ' interest in learning, and improve the willingness to learn a specific knowledge. However, students are less likely to think about the contents after the game, but keep playing down. Usually, it is until the final online evaluation begins to help the students to recall the contents of the previous. The process of the final review is important for the design. This will not only can immediately check whether the students have learned the knowledge, but also can let students have the chance through self-initiative to get the targeted knowledge. Based on the above consideration, this study proposes a new method for developing the game for learning by using the Role-Playing Game (RPG), chapters of textbooks and story development.

*Meng-Chien Yang, Department of Computer Sciences and Communication Engineering, Providence University. Email: mcyang2@pu.edu.tw

In the beginning, we design the guiding rules to lead students to play the game. Suppose that there is a specific topic prepared for learning. The students are told to play the game linked to this topic. After the game, the learners are given a review test for enhancing the learning results. Meanwhile, we using a set of animation films to assist the learners. These films can be traced back to the pre-test which is given before the students play the game. Then, after the review test students will watch the related film with the content of the learning topic. This will strengthen the concept of knowledge link with the students in this topic.

In this paper, we use the above framework to design the game for learning an important subject: Hamiltonian circuit of the graph theory in discrete mathematics. The Hamiltonian circuit problem is to find an existing loop to visiting each vertex once on the graph. The Hamiltonian circuit problem is an important problem and can be applied to many different technological applications. To develop a game for learning the Hamiltonian circuit problem is an important problem both in computer science education and mathematical education.

In recent years, developing math games have become the popular research topic. The importance of mathematical games in mathematics education cannot be ignored. A practical approach of using the game for teaching math in the country junior high school was reported [1]. In addition, the author pointed out that the mathematical game teaching method has the following advantages and functions:

(1) to stimulate students' interest and motivation to learn mathematics [2];

(2) to help students from the specific experience to establish basic mathematical concepts and skills;

(3) to cultivate students to solve the problem of flexible and deep reasoning thinking ability

In another paper [3], the experimental results and infield tests showed that the existence of mathematical games have considerable potential. Our paper follows these previous studies and will continue to explore the learning changes brought by this new design approach. Based on the hypothesis of the above research project, we hope our study of developing the game of the Hamiltonian circuit problem can help the students learning this problem. In our study, we had invited students who taking the course of the graph theory to play the game. The testing results showed our proposed framework can develop the effectiveness game to help student learning.

The remaining of this paper is organized as the follows. Section 2 discuss related studies of the game-based learning. Section 3 describe the design of our proposed framework. Section 4 is the description of the game development and system modules of the proposed game for the Hamiltonian circuit problem. Section 5 is the report of experimental results and discussion. The conclusions are put in Section 6.

2. Studies of Serious Game and Game-Based Learning

The research projects related to our study are the studies of the serious game. Based on the definition of the famous game design textbook, the serious games are those games for education and study, simulation, health and growth [4]. Serious game gives the players fun experiences and other explorations. Serious game design is built in the "fun" under the premise of the players in the process of the game, while achieving a sense of accomplishment, knowledge and training, and further use of the nature of media and entertainment to convey the effect of information and education [5][6]. A famous study argued that the practice of the serious game has the meaning of learning and is a systematic learning, so that students can learn from the experience of the game [7].

There are new studies that "video games" will be a very influential tool for learning [8][9].Game-based learning is not a regular learning style but a revolution for education. It can increase the learner's motivation to learn and help learners to learn effectively to improve learning motivation and effectiveness. Learners playing the game, easily immerse into the game and the learning scenarios because of game challenges, level design or script story into the immersive state. Therefore, game learning can follow the experience of "heart flow" [10], a process to obtain the knowledge, so that learners immersed in the game. The theory of flow suggests that when learners are integrated into the task, they produce a state of mind that is isolated from the outside world. If students are in such a state of mind, students will have the highest motivation to learn.

The study of the serious game development could have many different aspects. Recently a reviewing paper [11] described some essential features of the serious games for education. In this paper, the authors stated that two major focus of the serious game are learners' attitude and game core mechanics. Another important study about how to use the serious game for collegiate computer science students is to use the game with course materials [12]. The development of the serious game might need the guidelines for design the game. A recent study [6] proposed the guidelines and stated the key factors of the serious game development. Based on the above studies, several factor could influence the design and implementation of the learning game. First, the connection between the learning materials and the learning game should be tight. Second, the development of the game should follow a guideline that can make the game to become a useful learning materials. Then, the third, the learning outcome should be evaluated with different and various considerations and assumptions. Based on these assumptions, our study focus on designing a new framework for developing the learning game. In this framework, we developed a new learning outcome called learning reflexiveness which can help the learners to enhance the learners' attitudes in the game-based learning process. The framework can help designer to develop the serious learning games.

3. Methodology and Framework of the Game Design

The design of our framework was initiated with the concepts of the flow simulation proposed by Csikszentmihalyi [10], [13]. When human are involve the activities with flow based on the theory proposed by Csikszentmihalyi, people can concentrate on these activities and have the immediate feedback. The flow simulation had been used the development of the playground and video games. The activities with flow can bring self-consciousness of awarding the related information and context. Based on this theoretical finding, we attempt developing the learning process with the game. We describe our design by the question what is a

good educational learning game? This will provide scaffolding for our proposed framework.

A good educational game can motivate the learners through the design of the presumptive learning situations. The presumptive learning situations are a set of conditional factors for identifying the initial condition of each learning level. The learning situation and RPG game scenarios are the main settings of the game. Meanwhile, in the domain of the game, the story narration is a factor that attracts the learners from being immersed in it. We create stories and game rules for the game. The learners follow the guideline proposed by [14] in the symbol of the code, through the code to show the navigation keys to read the story. The strategy of game design for each level should be in a gradually better mode that can help the learner to immerse into the story of the game. Following previous study, the game prepares easy, moderate and difficulty challenges to enhance the learners' self-esteem [15]. The design is to design a learning situation during learning low-challenge tasks to learning high-challenge tasks. We developed the setting of the game by adopting the concept of flow in different working environments.

We had study the game development for several years. We analyzed and found the following three motivational factors that help learners to play the games [16].

1. Challenge

Each game level can be design with a particularly challenge. The challenge will entice learners to continue the game. The design of challenges is to describe a clear game goal, then an uncertain outcome, and a related feedback. The different challenges can be design with different learning materials.

2. Curiosity

The interface on the screen and sound control can cause learners interest. The setting of the game world can be designed as an environment that the players can explore new items.

3. Fantasy

Fantasy environment is an important factor. It stimulates the intrinsic motivation of learners in two ways, one is to satisfy the learner's inner emotion, and the second is to meet the learners' cognitive needs, and to promote the learners to grasp the relationships in the game.

With these three factors, we add a new factor into the game development. The new factor is the reflexiveness of the learner. Once the learner is playing the game, the outcome and the challenges can help the learner to have the learning attitude to adjust the learning paths toward the meaningful goal. The reflexiveness for learning had been mentioned in the studies of mathematics education.

3.1. The Importance of Reflexiveness for Mathematics Learning

Reflexiveness mathematics learning refers to the fact that it is an effective way to learn mathematics through the reflection of the process of math learning activities. It is not only a general review or repetition of mathematics, but a deep understanding of mathematics activities involved in the knowledge, methods, ideas, and strategies and so on. The general operational mathematics is based on "learning knowledge" as the main purpose, concerned about the current academic performance; reflexiveness mathematics is based on "learning to learn" for the purpose of focusing on the current academic performance and students' future development [17].

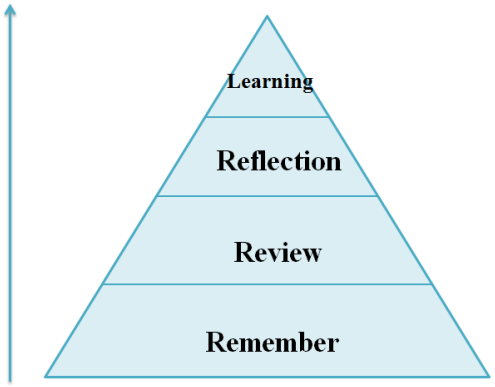


Figure 1: Reflectiveness of four learning levels

3.2. The reflexiveness in the proposed learning

Based on above discussions of the reflexiveness in mathematics learning, the reflexiveness can be applied to the other learning subject. The reflexiveness in the game-based learning can be illustrated as four different levels shown on Figure 1. The bottom level is the learning for remembering. This can be the basic of reflexiveness and are used as the fundamentals for accumulating knowledge. The second level is the learning with the reviewing content. The function of the second level is to enhance the effectiveness of the learning. The third is reflection which are the main level of our design. This level is to find the correct ways of learning the targeted subjects or courses. The final level is the effective learning process that the learners can successfully learn this subject.

We used the proposed framework to develop a game for learning the Hamiltonian circuit problem. The Hamiltonian circuit is an important graph in the graphic theory. The Hamiltonian circuit problem can be taught in the collegiate courses such as discrete mathematics, graph theory, combinatorial mathematics. The Hamiltonian circuit problem is an important concept in many computer science applications like logic design, routing problem. In next section, we describe our game that following the proposed framework.

4. Game Design and Game Play

In this section, the system design of the game is illustrated. At the beginning, the whole structure the game is to design with an RPG game structure combined with the fairy tale. Figure 2 shows the main menu of the game. The game is described as several chapters with different concepts of the Hamilton circuit. At the beginning of each chapter the learners will receive guidance rules. The guiding rules are put in a short RPG game and with a short introduction video clip. When the learner plays the RPG game, he can get the game rules. Using these rules, the learners can play and find the Hamilton circuit in a given graph (Figure 3). If the learner can successfully find the Hamilton circuit in the graph, the learner

can be guided to complete the chapter with more similar graphs. When the learners finish the chapter of the game, there is a chapter reviewing quiz which helps the learners to recall the process of gameplay. The learners can get an immediate feedback here and understand the correct answers which are the key terms of this chapter.

Figure 2 Main menu of the game

The reflexiveness setting in the whole scenario of the gameplay is a cyclic framework shown in Figure 4. In this framework, the learner can find the learning content in four different styles. The guide rules and the RPG game are used to attract the learner’ s attention. The game process is the main challenge and help the learner to understand the Hamilton circuit. The review test can help the learner to enhance the learning results. The teaching video can keep the learner’s interest and go to next chapter which is more difficult Hamiltonian graph.

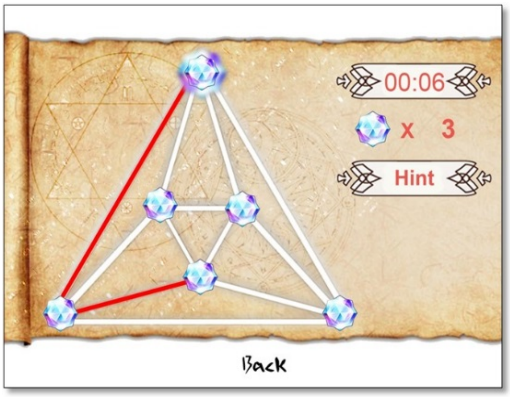


Figure 3: Gameplay for finding the Hamiltonian circuit

In addition, the reflective learning will affect the learners in two factors of each cycle. First, in the review of the test (Figure 5), learners need to recall the game experiences for answering these questions correctly. Second, the animated films with chapter concept guide, learners can link the game experience and answer process. An interesting design here is that the outcome of the story will vary depending on the score of the learner's game, which gives the learner the motivation to repeat the game.

Reflective learning should help the learners to get attention to the guidance and encouragement in the learning process. When the learners get the guide rules after the game, and in the review of the test to reflect on the game process and guide the rules of the tips, organize the answer to answer questions. When learners watch the teaching film for the second time of reflective activities, the learners will have reflective experience based on the pasting activities.

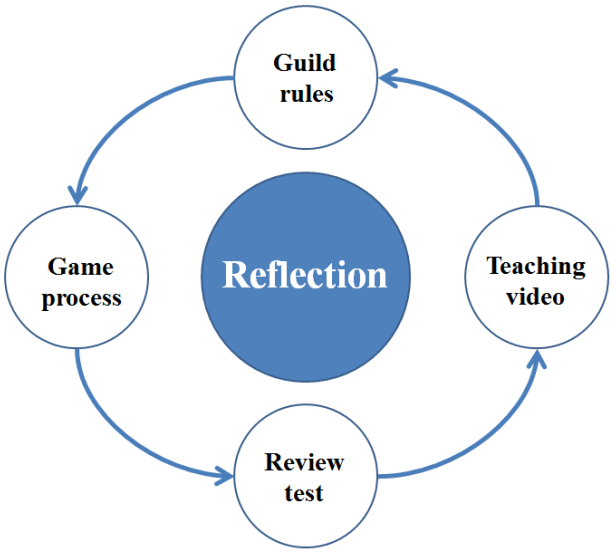


Figure 4: Game cycle that includes reflective learning.

The game is implemented with this system design. In the game, six cycles for the easy to complicate knowledge of the Hamilton circuit are implemented. After the review test, the learners can watch an animated film as the enhancement for learning reflection.

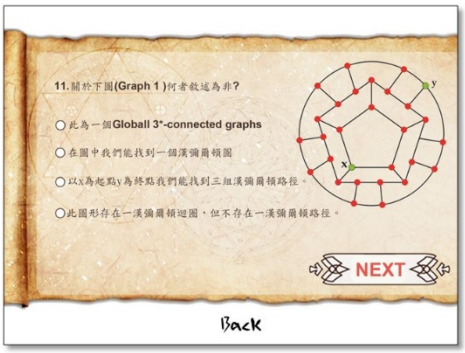


Figure 5: Screen shot of the online reviewing test

5. Implementation and Evaluation

The learning game was implemented and evaluated by a user–based evaluation and an expert-based working through evaluation. The approach was similar to a previous study of the educational serious game development [12]. A group of computer science major college students who taking the graph theory course were invited to participate the user-based evaluation. Our game was installed in a tablet computer and each participant was given 10 minutes to play the first three chapters of the game and then answered the quiz after the game. If the student could finish these three chapters successfully, the student would be given a big reviewing test and were asked to finish the remaining three chapters of the game.

The total number of students participating this experiment were 34. When the students finished the first three chapters of the

game, they were given a writing test for evaluating the basic concepts of the Hamiltonian circuit. When they finish the final chapter, the sixth chapter, the students were given a writing test with the complex knowledge of Hamiltonian circuits.

The experimental results showed 19 students had finished the first three chapters of the game and passed the basic test. Then, 15 students had finished the whole six chapters of the game and passed the advance test. Based on the evaluating, the developed game can be useful as a tool for teaching Hamiltonian circuit.

The second evaluation process is the expert walk-through test. We invited 8 computer major graduate students who had already understood the Hamiltonian circuit. When these students played the game, they were asked to find the shortcoming of the usability of the interface and evaluating the gameplay. Most students agreed that the game can help the beginner to learn the Hamiltonian circuit. Several mentioned the advanced part, the last three chapter should be improved with more illustration figures to help the learners play the games. Based on the suggestions, we are certain that the proposed framework can foster a comprehensive and effective learning.

6. Conclusion

In this paper, a framework for developing the learning games for advance computer science concept is proposed. In this framework, the learning process is arranged as a cycle with playing game, answering the review questions, watching the animation clips and being given the guiding rules. The framework is developed based on the concept of the flow and the fundamentals of the game theory. In addition, the reflexive review of the mathematical education is put in the framework and we coin a new learning attitude, reflexiveness. The reflexiveness design in the learning game can help the learners achieve the learning goal via playing.

We used this framework for developing the game for learning the comprehensive knowledge of Hamiltonian circuit. The results of evaluation showed that the game can help these participants to learn the Hamiltonian circuit. The game can be a useful teaching tool for courses related to graph theory and computer science. The game can be used for other purpose such that training course for novices to learn this concept at the IC design company.

The project has two successful results. First, we have develop a framework that can be useful as a guideline for other research studies who focus on developing the serious games for certain subject. Second, the game for teaching the Hamiltonian circuit can be useful as the teaching tool. In the next studies, we will explore and find more evidences that proposed framework is effective for the developing the serious learning game.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research was supported in part by Providence University 2017 grant for graduate student.

References

[1] C. W. Rau, Game-Based Teaching for Elementary Math Courses.Wu-Nan Culture Publish Co. 1996.

[2] E.R. Lai, Motivation: A Literature Review. Pearson. Research Report. Pearson's publications 2011.

[3] C. W. Lai, & P. J. Lin, Immersion of Mathematic Games for high school math function courses. Master Thesis, National Hsinchu Educational University, 2011.

[4] E. Adams, Fundamental of Game Design 3rd, New Riders, 2015.

[5] A. Singhal, & E. M. Rogers, "A theoretical agenda for entertainment—education." *Communication theory*, *12*(2), 117-135, 2002.

[6] E., Tsekleves, J. Cosmas, & A. Aggoun, "Benefits, barriers and guideline recommendations for the implementation of serious games in education for stakeholders and policymakers." *British Journal of Educational Technology*, *47*(1), 164-183, 2016.

[7] A. Derryberry, "Serious Games: Online Games for Learning," Adobe White Paper 2007.

[8] K. Humphrey, "The application of a serious, non-digital escape game learning experience in higher education," *Sport & Exercise Psychology Review*, *13*(2), 48-54, 2017.

[9] F. W. Kron, C. L. Gjerde, A. Sen, M. D. Fetters, "Medical student attitudes toward video games and related new media technologies in medical education." *BMC Medical Education*, *10*(1), 50, 1-11, 2010.

[10] M. Csikszentmihalyi , Creativity: Flow and the Psychology of Discovery and Invention, New York: Harper Perennial, 1996

[11] P. Lameras, S. Arnab, I. Dunwell, C. Stewart, S. Clarke, & P. Petridis, "Essential features of serious games design in higher education: Linking learning attributes to game mechanics." *British Journal of Educational Technology*, *48*(4), 972-994, 2017.

[12] A. Schäfer, J. Holz, T., Leonhardt, U. Schroeder, P. Brauner, & M. Ziefle, "From boring to scoring – a collaborative serious game for learning and practicing mathematical logic for computer science education." *Computer Science Education*, *23*(2), 87-111, 2013.

[13] J. Nakamura, & M. Csikszentmihalyi, "The concept of flow." In *Flow and the foundations of positive psychology* (pp. 239-263). Springer Netherlands, 2014.

[14] M. P. Chen," Learner's attitude and evaluation for programming language courses", Journal of National Normal University, 52, 1-21, 2007.

[15] T. W. Malone, & M. R. Lepper, "Making learning fun: A taxonomy of intrinsic motivations for learning." *Aptitude, learning, and instruction*, *3*, 223-253, 1987.

[16] M. Papastergious, "Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation", Computers and Education, 52(1), 1-12.2009.

[17] R. B. Tu, "Discussion of Reflexiveness for Mathematic Education." Journal of Mathematic Education, 4(3), 2000.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 357-365 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Theoretical Investigation of Combined Use of PSO, Tabu Search and Lagrangian Relaxation methods to solve the Unit Commitment Problem

Sahbi Marrouchi*, Nesrine Amor, Moez Ben Hessine, Souad Chebbi

Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE), National Superior School of Engineers of Tunis (ENSIT), University of Tunis, 5 Street Taha Hussein Montfleurie, 1008 Tunis, Tunisia

ARTICLE INFO

Article history:
Received: 30 November, 2017
Accepted: 15 January, 2018
Online: 02 February, 2018

Keywords:
Unit commitment
Optimization methods
Scheduling
TS-PSO-LR

ABSTRACT

Solving the Unit Commitment problem (UCP) optimizes the combination of production units operations and determines the appropriate operational scheduling of each production units to satisfy the expected consumption which varies from one day to one month. Besides, each production unit is conducted to constraints that render this problem complex, combinatorial and nonlinear. In this paper, we proposed a new strategy based on the combination three optimization methods: Tabu search, Particle swarm optimization and Lagrangian relaxation methods in order to develop a proper unit commitment scheduling of the production units while reducing the production cost during a definite period. The proposed strategy has been implemented on a the IEEE 9 bus test system containing 3 production unit and the results were promising compared to strategies based on meta-heuristic and deterministic methods.

1. Introduction

The main role of energy management is to ensure the production of active power in order to respond to the demand growth among a very little fuel cost [1,2]. Solving the Unit Commitment Problem (UCP) is very essential in electrical network planning. It can both optimize the daily operational planning of networks and reduce the total production cost among the improvement of the operating state of each unit leading to obtain the best unit commitment scheduling helping to respond to the power demand. Operations scheduling production units or Unit Commitment (UC) improve operational planning of the electrical grid while ensuring continuity of service [1-6]. The main purpose of solving the Unit Commitment problem is to schedule production units to respond to the consumed power with the minimization of the total production cost. The optimal planning [7-9] involves ensuring a better use of available generators subject to various constraints and guaranteeing the transfer of electrical energy from generating stations to the load. UC must satisfy the load demand, storage capability, minimum downtime startup and safety limits for each production unit.

The production scheduling comprises determining startup and each generation level for each unit in a given planning period [10-13]. Therefore, a study of literature [14,15] on methods which focus on unit commitment (UC) problem resolution shows that various optimization methods have examined this subject. Furthermore, Sasaki et al. demonstrated the possibility to use artificial neural network (ANN) to solve the UCP in which a large number of inequality constraints is processed. They have used the ANN to schedule generators and the dynamic programming to solve the load flow problem. The adopted strategy was compared to Lagrangian Relaxation (LR) and dynamic programming (DP) methods and the results offered a faster and cheaper solution compared to the LR and DP but it suffers from digital convergence because of the learning process. Certain works [16, 17] proposed a strategy based on tabu search method. They introduced new rules to generate an initial solution feasible to solve the Unit Commitment problem. This strategy consists on dividing the problem into two problems: the first combinatorial optimization problem is solved using tabu search algorithm and the second is a problem of nonlinear programming solved through the quadratic

*Corresponding Author: Sahbi Marrouchi, National Superior School of Engineers of Tunis (ENSIT), University of Tunis, Tunis, Tunisia
Email: sahbimarrouchi@yahoo.fr

programming routine. The structure resolution through Tabu search method is similar to that used by simulated annealing [18] even though TS is provided with a simplified configuration, so it is easy to pass from one optimization to the other. Indeed, the main advantage of the adopted strategy is to extend the search space provided for the best optimal solutions which are stored in the tabu list. This method has provided a lower production cost solution, but it's slower compared to the Lagrangian relaxation. However, T. Logenthiran et al. [19] have proposed a new approach based on particle swarm optimization (PSO) algorithm for solving the unit commitment problem. They presented three versions of particle swarm: binary particle swarm optimization (BPSO), improved binary particle swarm optimization (IBPSO) and combined use of particle swarm optimization and Lagrangian relaxation programming (LR-PSO) .The numerical results show that LR-PSO method has provided a lower production cost solution compared to LR, BPSO and IBPSO especially when the number of units important. Whereas, if the number of units is small, BPSO is taken as the best method since it has the lowest production cost compared to other algorithms. Other works, [20, 21] presented new approaches based on artificial intelligence to solve the UCP. The adopted approach combines two methods: tabu search and neural networks (ANN-TS) in order to get an optimal unit commitment scheduling allowing a minimal production cost in accordance to the constraints of the studied system. Artificial Neural networks provide a fast convergence to optimal solutions but it takes a lot of memory space because of the great number of constraints.

Cheng et al. [22] proposed a hybrid method based on the integration of the genetic algorithm in Lagrangian Relaxation Programming (LR-GA) to solve the problem of the planning of the operations of the production units. This integration consists in improving the Lagrange multipliers using the operators of the genetic algorithm to find a fast and an effective cost solution respecting all the constraints of the system. The implementation of this method requires two steps ; the first is to look for the minimum constraints of the Lagrange function under the multipliers constraint through dynamic programming. The second step consists on maximizing Lagrange's function while respecting the multiplier adjusted by the genetic algorithms. The experimental result of this method provides a faster and cheaper solution compared to the Lagrangian relaxation method (LR) and the tabu search method (TS). However, C. Christober et al. [23] presented a hybrid method combining the evolutionary programming the tabu search methods (EP-TS). The first has the advantage of a good convergence property, a significant acceleration based on the principle of traditional genetic algorithm and a high quality of solutions, but its major disadvantage is related to the dimensioning of the Unit Commitment problem. Tabu search method improves the status by avoiding imprisonment in the local minimum. The best solution is chosen by evolutionary strategy. Thus, the effort has been made to combine these two methods whose purpose is to meet the requirements of the generators commitment problem. The numerical results demonstrate that this method is efficient and accurate in terms of calculation time and minimization of the total production cost compared to the following methods: simulated annealing, taboo search, dynamic programming, evolutionary programming, Lagrange programming and LR- GA. By contrast, Kumar et al. [24] presented a hybrid method combining dynamic programming with Hopfield Neural Networks (DP-HNN). The proposed process consists on two steps: use of Hopfield neural networks from direct computation to generate the economic distribution (ED) and use of dynamic programming (DP) to plan generators. This approach provides a poor final solution in total of production cost over other methods, but is faster than LR, DP, GA, LR-GA and AS-GA. In addition, C. Asir et al. [25] presented a new approach based on artificial intelligence to solve the problem of allocation of units. This combines two methods: tabu research and Artificial Neural Networks (ANN-TS) and this in order to have an optimal solution that solves the problem of planning power with a minimum of total production cost with respecting all the constraints of the specified system. Neural networks provide a fast convergence solution but the programming of the algorithm takes up a lot of memory space because of the constraints of the problem. The tabu search is characterized by the flexibility of its memory and it is able to find good solutions. This algorithm gave a faster and cheaper result compared to tabu search, dynamic programming, Artificial neural networks, Lagrange programming, LR-GA, TS-GA and EP-TS methods. In addition, C. Asir [26] has developed another strategy which consists in integrating the tabu search with the genetic algorithm. The purpose of this technique is to find the planning of the production. This strategy depends on the exploitation of the total cost which can be minimized when it is subjected to a set of constraints. Tabu search can find good solutions using the tabu list while genetic algorithm is used to generate new solutions using its operators. The results demonstrate that this technique provides a low cost but slow solutions compared to ANN-TS. Alma et al. [27] proposed a hybrid approach combining fuzzy logic and the genetic algorithm (FL-AG) to solve the Unit Commitment problem. Genetic algorithm uses the actual coded chromosomes in contradiction with the most commonly used scheme which is binary coding. This method uses a strict priority order list in the genetic algorithm that generates different solutions. This list serves to reduce the size of the search space of the problem, while fuzzy optimization guides the entire search process in an uncertain environment (varying from load demand, renewable energy sources etc ...).The results of this technique offer a good final solution compared to dynamic programming and to the genetic algorithm.

According to our study, we thought to validate an approach to apprehend the whole unit commitment problem. To achieve this objective, our strategy for solving the Unit Commitment Problem is based on the combination of three stochastic optimization methods that are the Particle Swarm Optimization (PSO), the Tabu Search (TS) and Lagrangian Relaxation (LR) method in order to develop a proper unit commitment scheduling of the production units to minimize the production cost.

2. Notation

The notation used throughout the paper is stated below.

a_i, b_i, c_i : Coefficients of the production cost,

P_{ih} : Active power generated by the i^{th} unit h^{th} hour, $i = 1,2,3,....,N_g$ and $h = 1,2,3,....,H$

U_{ih} : On/Off status of the i^{th} production unit at the h^{th} hour, $U_{ih} = 0$ for the off state of one generating unit and $U_{ih} = 1$ for the operating status of one generating unit,

HSC_i : Hot start-up cost of the i^{th} unit,

CSC_i : Cold start-up cost of the i^{th} unit,

P_{rh} : System spinning reserve at the h^{th} hour,

P_{dh} : Amount of the consumed power at the h^{th} hour,

P_{Lh} : Total active losses at the h^{th} hour,

P_i^{min} : Minimum and maximum power produced by one generator,

P_i^{max} : Maximum power produced by one generator,

MUT_i : Continuously on-time of unit i .

MDT_i : Continuously down-time of unit i .

τ_i^{OFF} : Continuously off-time of unit i ,

SC_i : Cold start time of unit i .

N_g : Number of generating units,

H : Time horizon for UC (h).

3. Problem Formulation

Many works have been based on an analytical statement of the unit commitment problem [2,3,11, 18, 22]. We present in this paper a mathematical model of the unit commitment problem with limited security. This model is a mixed linear and constrained which has been adapted in several works [3, 9, 13, 15].

$$Min \left[F_T(P_{ih},U_{ih}) = \sum_{i=1}^{N_g} \sum_{h=1}^{H} [a_i P_{ih}^2 + b_i P_{ih} + c_i + ST_i(1-U_{i(h-1)})]U_{ih} \right] \quad (1)$$

Where;

ST_i: The starting cost of the i^{th} unit defined by:

$$ST_i = \begin{cases} HSC_i \ \ if \ \ MDT_i \le \tau_i^{OFF} \le MDT_i + SC_i \\ CSC_i \ \ if \ \ \tau_i^{OFF} \succ MDT_i + SC_i \end{cases} \quad (2)$$

The minimisation of the objective function is provided with the following constraints:

- System Constraints
  - Power balance constraints

$$\sum_{i=1}^{N_g} P_{ih} U_{ih} = P_{dh} \quad (3)$$

  - Spinning reserve constraints

$$P_{dh} + P_{rh} - \sum_{i=1}^{N_g} U_{ih} P_{ih} \le 0 \quad (4)$$

- Unit Constraints
  - Genration limits

$$P_i^{min}.U_i \le P_{ih}.U_i \le P_i^{max}.U_i \quad (5)$$

  - Minimum up-time constraint

$$U_{ih} = 1 \quad for \sum_{t=h-up_i}^{h-1} U_{ih} \le MUT_i \quad (6)$$

  - Minimum down-time constraint

$$U_{ih} = 0 \quad for \sum_{t=h-down_i}^{h-1} U_{ih} \le MDT_i \quad (7)$$

4. Methodology of resolution

In this paper, four optimization methods are available to solve the unit commitment problem; the first one uses the Particle Swarm Optimization (PSO). This strategy takes into account the advantage of PSO method for solving complex and nonlinear problems. The second method relies on the use of the Tabu Search approach (TS). The use of the TS approach is depicted to the flexibility of storage great memory of optimal solutions offered by this method. The third strategy shows the advantage of the Lagrangian relaxation providing the best convergence speed. Our strategy for solving the Unit Commitment Problem is based on the combination of three optimization methods that are the Particle Swarm Optimization (PSO), the Tabu Search (TS) and Lagrangian Relaxation (LR) method to find a good On / Off states scheduling of each production unit over a period of time leading to obtain a good production cost.

4.1. Particle Swarm Optimization

Particle swarm optimization provides a population based search procedure in which individuals called particles change their positions with time. This method is able to generate high quality of solutions within shorter calculation time and stable convergence characteristic than other stochastic optimization methods. The PSO model consists of a swarm of particles moving, figure 1, in a definite dimensional real-valued space of possible problem solutions [9,28,29].

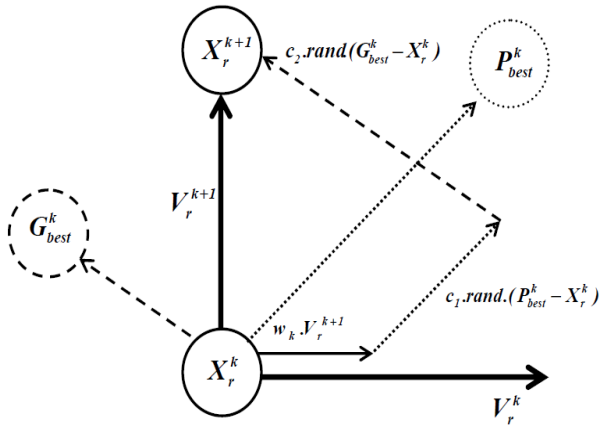


Figure 1. PSO research process

Every particle has a position $X_i = (x_i^1, x_i^2, ..., x_i^l)$ and a flight velocity $V_i = (v_i^1, v_i^2, ..., v_i^l)$. Indeed, each particle has its own best positions $P_{ibest} = (P_{ibest}^1, P_{ibest}^2, ..., P_{ibest}^l)$ and a global best position $G_{best} = (G_{best}^1, G_{best}^2, ..., G_{best}^l)$. Each time step is characterized by the update of the velocity and the particle is moved to a new position which is the sum of the previous position and the new velocity as shown in the following expression:

$$X_r^{k+1} = X_r^k + V_r^{k+1} \quad (8)$$

The update of the velocity from one particle to another is given by:

$$V_r^{k+1} = w_k.V_r^{k+1} + c_1.rand.(P_{best}^k - X_r^k) + c_2.rand.(G_{best}^k - X_r^k) \quad (9)$$

Where, c_1 et c_2 are acceleration constant, *rand* is a uniform random value between $[0,1]$, X_r^k and V_r^k are respectively the position and the velocity of one particle i at iteration k. w_k is the inertia weight factor defined by the following equation:

$$w_k = w_{max} - \frac{w_{max} - w_{min}}{k_{max}}.k \quad (10)$$

Where, w_{max} and w_{min} are the maximum and the minimum inertia weight factors respectively and k_{max} is the maximum number of iterations.

4.2. Tabu search

Tabu search uses a local or neighborhood search procedure to iteratively move from a solution X to a solution X' in the neighborhood of X, until some stopping criterion has been satisfied. To explore regions of the search space that would be left unexplored by the local search procedure, TS modifies the neighborhood structure of each solution as the search progresses [16,17]. The search for the optimal solution corresponding to minimal production cost consists on repeating an iterative process until reaching a stop criterion so as to find one solution neighbor to the optimal one as shown to the following equation:

$$P_{ibest} = P_{best} + \left[(b_i + 2.c_i.P_{best}).\left(\frac{P_i^{max} - P_i^{min}}{2} \right) \right] \quad (11)$$

The new neighborhood solutions $N^*(X)$ are determined through the use of memory structures. The search then progresses by iteratively moving from a solution X to a solution X' in $N^*(X)$. To determine the solutions admitted to $N^*(X)$, a tabu list (TL) memory is used, which is a short-term memory containing the solutions that have been visited in the recent past as less than the maximum number of iterations.

4.3. Lagrangian Relaxation

The Lagrangian relaxation solves the Unit commitment problem by relaxing or temporarily ignoring the constraints, power balance and spinning reserve requirements [13,15,30]. Therefore, to transform the complex nonlinear constrained problem into a linear unconstrained problem, we have considered the following Lagrangian function:

$$L(P_{ih}, U_i, \lambda_i) = \sum_{i=1}^{N_g}\sum_{h=1}^{H}[a_i P_{ih}^2 + b_i P_{ih} + c_i + ST_i(1 - U_{i(h-1)})]U_{ih} + \lambda_i.(P_d - \sum_{i=1}^{N_g} P_i U_{ih}) \quad (12)$$

Where, λ_i is the Lagrangian coefficient.

To establish our strategy, we have considered the partial derivatives of the Lagrangian function (12) with respect to each of the controllable variables equal to zero.

$$\frac{\partial L}{\partial P_{ih}} = \frac{\partial\left[[a_i P_{ih}^2 + b_i P_{ih} + c_i + ST_i(1 - U_{i(h-1)})]U_{ih}\right]}{\partial P_{ih}} - \lambda_i\left(\frac{\partial P_{dh}}{\partial P_{ih}} - U_{ih}\right) = 0 \quad (13)$$

$$\frac{\partial L}{\partial \lambda_i} = P_{dh} - \sum_{i=1}^{N_g} P_i U_{ih} = 0 \quad (14)$$

Equations (13) and (14) represent the optimality conditions necessary to solve equation systems (1) and (3) without using inequality constraints (equations (4) and (5)). Equation (13) can be written as follows:

$$\lambda_i = \frac{\frac{\partial\left[a_i P_{ih}^2 + b_i P_{ih} + c_i + ST_i(1 - U_{i(h-1)})]U_{ih}\right]}{\partial P_{ih}}}{\frac{\partial P_{dh}}{\partial P_{ih}} - U_{ih}} \quad ; i = 1,\ldots,N_G; h = 1,\ldots,H \quad (15)$$

4.4. Proposed Strategy

The process of the Unit Commitment problem resolution by the combined use of Tabu search, Particle swarm optimization and Lagrangian Relaxation (TS-PSO-LR) methods is carried out according to the flowchart in Figure 2.

The proposed strategy not only helps to reach the optimal solution as quickly as possible using the speed of the Lagrangian relaxation but also to proceed through PSO method to search effective solutions corresponding to a minimum production cost and this is obtained through a specific determination of the new velocity and then the next best position corresponding to the best amount of generated power produced by each unit when it's in the ON state.

In the proposed method, it's notable that P_{ibest} representing the best information of each particle and the history of each generated power P_{gi} of each production unit is preserved in the list $P_{best}List$. Herein, in spite of the possibility of the PSO method with the solution better than G_{best}^k around P_{best}^k, there is the possibility not to be searched enough that's why we have thought to use the history of P_{ibest}^l in $P_{best}List$.

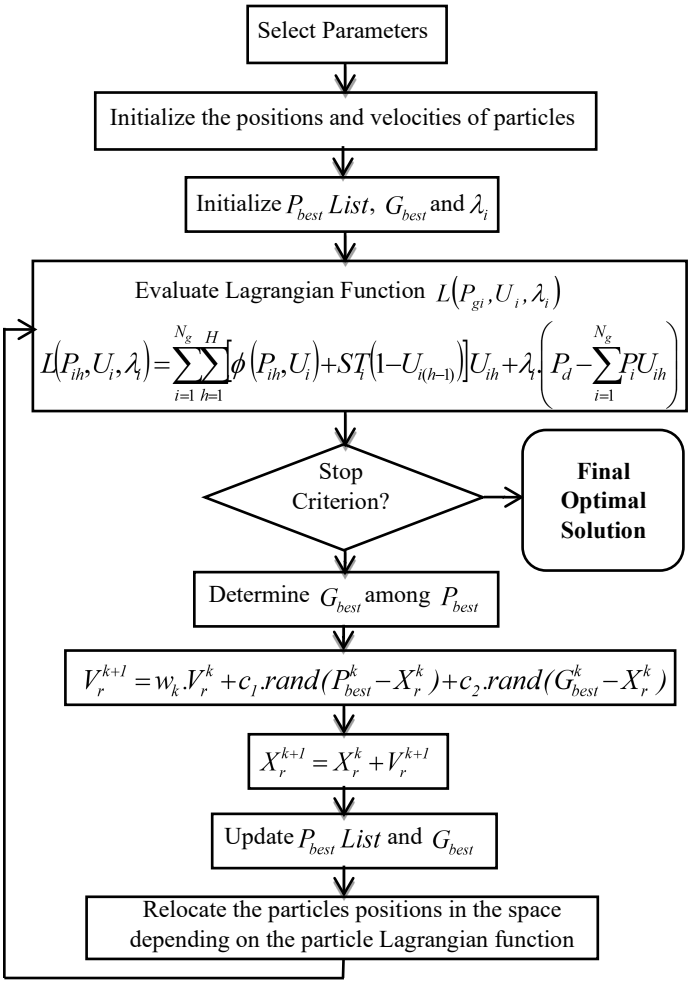


Figure 2. Flowchart of solving the unit commitment problem via Tabu Search, Particle Swarm Optimization and Lagrangian Relaxation

Whenever the particles lose the searching ability when the velocity V_r^k of one particle is very small, the TS-PSO-LR algorithm adapts the other $(P_{ibest}^2,...,P_{ibest}^l)$ instead of P_{ibest}^1 to update equation of velocity. This action increases the searching ability and helps to find more optimal solutions enabling a minimal production cost while considering a best unit commitment scheduling.

The proposed TS-PSO-LR strategy differs from other evolutionary computing techniques in providing an acceptable solution within a relatively short time and is likely to lead the search towards the most promising solution area. A step-by-step TS-PSO-LR for the optimization of the UC problem is outlined as follows:

Step 1: Initialization data for each unit production

Step 2: Create tabu lists $P_{best}List$

Step 3: Generate the position and the velocity (X_r^k, V_r^k) of each particle according to equations (8) and (9).

Step 4: Evaluation of each particle at an initial power value selected from the limit margins.

Step 5: Calculate the production cost function of each particle for each production unit,

Step 6: Calculate the Lagrangian coefficient λ according to the expression (15),

Step 7: Calculate the objective function of each particle according to equation (1),

Step 8: Save the best results in the tabu list $P_{best}List$

Step 9: If the stop criterion is satisfactory, the found values are those which corresponds to the desired optimal solution otherwise we return to step 3.

5. Simulation And Results

In order to test the performance of the optimization proposed method; the strategy has been applied to an IEEE electrical network 9 buses [13,15,31], having 3 generators, over a period of 48 hours. The characteristics of the different production units are given in Table 1.

Table 1: Characteristics of production units

U	P_{gimax} (MW)	P_{gimin} (MW)	a	b	c	MUT	MDT	HSC_i ($)	CSC_i ($)
1	582	110	0.0756	30.36	582	8	8	4500	9000
2	330	74	0.00031	17.26	970	8	8	5000	10000
3	115	25	0.00211	16.5	680	5	5	560	1120

In this paper, we have considered 48 successive periods in order to establish the temporal evolution of the power demand (Table 2).

Table 2: Amount of load required

H	Load (MW)	H	Load (MW)	H	Load (MW)	H	Load (MW)
1	353.2	**13**	993.2	**25**	833.2	**37**	682.2
2	378.5	**14**	913.2	**26**	813.2	**38**	715.2
3	463.2	**15**	853.2	**27**	763.2	**39**	773.2
4	573.2	**16**	725.2	**28**	713.2	**40**	843.2
5	628.2	**17**	613.2	**29**	626.2	**41**	883.2
6	693.2	**18**	580.2	**30**	547.2	**42**	911.2
7	713.2	**19**	673.2	**31**	503.2	**43**	945.2
8	753.2	**20**	730.2	**32**	473.2	**44**	960.2
9	843.2	**21**	835.2	**33**	433.2	**45**	1001.2
10	925.2	**22**	945.2	**34**	533.2	**46**	1003.2
11	963.2	**23**	1007.2	**35**	583.2	**47**	925.2
12	1013.2	**24**	893.2	**36**	627.2	**48**	823.2

Table 3: Comparative table of the different methods used to solve the UC problem

H	P_{dh} (MW)	Production Cost ($)				Unit Commitment scheduling			
		LR	PSO	TS	PSO-TS-LR	LR	PSO	TS	PSO-TS-LR
1	353.2	23000	14198	14012	**3424**	111	111	111	**111**
2	378.5	25000	12981	15396	**13504**	111	111	111	**110**
3	463.2	29000	19729	20582	**13261**	111	111	111	**110**
4	573.2	375000	20320	28638	**15693**	111	110	111	**110**
5	628.2	845000	25275	33228	**19913**	110	110	111	**110**
6	693.2	1278000	32907	39192	**28268**	110	110	111	**110**
7	713.2	1432000	33600	41136	**19099**	110	110	111	**111**
8	753.2	1770000	43686	45172	**25722**	110	110	111	**111**
9	843.2	2688000	44498	50251	**32251**	110	110	111	**111**
10	925.2	7267000	46943	51667	**41758**	111	111	111	**111**
11	963.2	8441000	52566	52325	**46546**	111	111	111	**111**
12	1013.2	10129000	53198	53197	**53172**	111	111	111	**111**
13	993.2	9434000	53198	53197	**50477**	111	111	111	**111**
14	913.2	6915000	51897	49806	**40681**	111	111	111	**111**
15	853.2	2804000	37910	47580	**33298**	110	111	110	**111**
16	725.2	1529000	25883	46852	**29262**	110	111	110	**110**
17	613.2	758000	21020	46852	**25554**	110	111	110	**110**
18	580.2	584000	16014	46852	**14909**	110	111	110	**110**
19	673.2	1134000	29433	47666	**23751**	110	111	110	**110**
20	730.2	1571000	32618	49493	**33494**	110	111	110	**110**
21	835.2	2597000	38861	53197	**42276**	110	111	110	**110**
22	945.2	7873000	52140	53197	**50791**	111	111	111	**111**
23	1007.2	9918000	53198	53197	**52967**	111	111	111	**111**
24	893.2	6349000	51158	53197	**37898**	111	111	111	**111**
25	833.2	4797000	35703	53197	**45178**	111	111	111	**111**
26	813.2	4327000	36113	48240	**28933**	111	111	111	**111**
27	763.2	1861000	28794	47371	**24434**	111	111	110	**111**
28	713.2	1432000	25401	46852	**29117**	111	111	110	**111**
29	626.2	833000	46373	43893	**38959**	110	110	110	**101**
30	547.2	432000	20724	43893	**31753**	110	110	100	**101**
31	503.2	260000	24171	43893	**24453**	110	110	100	**101**
32	473.2	30000	21804	43893	**21221**	110	110	100	**101**
33	433.2	28000	25661	43893	**17202**	110	110	100	**101**
34	533.2	33000	29609	46499	**30381**	111	110	100	**101**
35	583.2	485000	45736	46499	**33614**	111	110	101	**101**
36	627.2	1017000	31241	46499	**39093**	111	110	101	**101**
37	682.2	1198000	46649	53197	**22191**	111	110	101	**111**
38	715.2	1448000	36247	53197	**26763**	111	111	111	**111**
39	773.2	1955000	35234	53197	**29056**	110	111	111	**111**
40	843.2	2688000	39180	53197	**32182**	110	111	111	**111**
41	883.2	3169000	42243	53197	**36731**	110	111	111	**111**
42	911.2	6857000	51986	53197	**40051**	110	111	111	**111**
43	945.2	7873000	52083	53197	**52034**	110	111	111	**111**
44	960.2	8345000	50398	53197	**51313**	110	111	111	**111**
45	1001.2	9709000	53198	53197	**52755**	110	111	111	**111**
46	1003.2	9778000	53198	53197	**52831**	111	111	111	**111**
47	925.2	7267000	51916	53197	**41756**	111	111	111	**111**
48	823.2	2464000	34368	49284	**30006**	111	111	110	**111**

Simulation results shown in Table 3 have proved that the adopted optimization methods have helped to establish an appropriate On/Off scheduling operating states of the production units while respecting the time constraints.

Nevertheless, through these methods we have arrived to reach an optimal production cost. Based on Table IV, the production cost found by the hybrid method based on the combination between Tabu search, Particle Swarm Optimization and Lagrangian Relaxation methods (TS-PSO-LR) method among 48 hours is about 1.5800e+06 $ lower compared to that obtained through Tabu search (PC = 2.2350e+06 $) and through PSO (PC= 1.7813e+06 $) or through Lagrangian Relaxation (PC= 1.6405e+08$). This result shows the best performances of the adopted strategy in minimizing the production cost and proves that we can get promising results through hybridization.

Table 4: Production Cost and Time required to converge for each optimized method

	TS	PSO	LR	TS-PSO-LR
Production Cost ($)	2.2350e+06	1.7813e+06	1.6405e+08	1.5800e+06
Time (s)	1.536 s	2.432s	4539s	127.082 s

Furthermore, concerning the resolution time, TS and PSO methods has presented the best time of convergence to an optimal solution compared to our strategy which requires 127.082 s to reach the global optimum. Besides, through Lagrangian Relaxation method, the unit commitment problem requires a lot of time to converge and this is explained by the complexity of the problem.

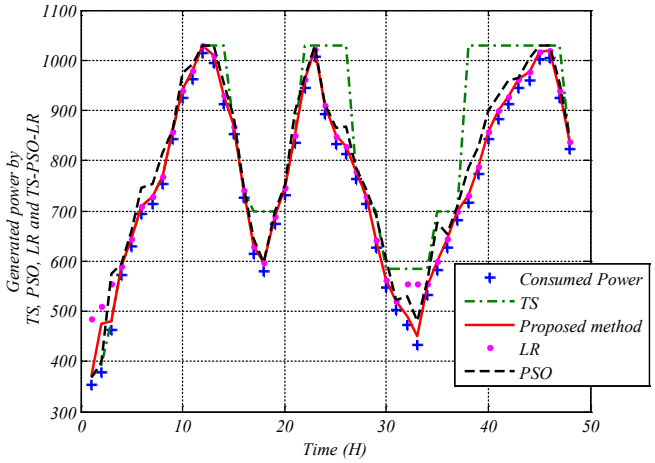


Figure 3. Generated power by TS, PSO, LR and TS-PSO-LR methods

It is interesting to note that the generated powers follow the optimum power quantities provided by the proposed optimization algorithm and the other optimization method. This demonstrates the high performance of the control algorithms adopted for the supervision of the system studied and proves the efficiency of the regulation loops for the different production units. In addition, the strategy adopts a permit to obtain sufficient and rapid planning in terms of convergence. Indeed, with the particular attention we offer you the considerable choice of input variables to particle swarm optimization method, we have managed via the proposed strategy to optimize the optimal solutions able to reduce the total cost of production.

Based on the results set out in Table 4, we find that our strategy has solved the Unit Commitment problem while addressing a planning of on-off states of production units. Planning that has complied with the constraints of each unit (minimum start-up MUT_i and shut-down times MDT_i). Moreover, we note that the power produced by the most powerful machine (615 MVA) remains unchanged throughout the 48 hours while the other production units vary to produce the amount of power demanded by the network.

Thanks to the simulation results, we can note that the Unit Commitment scheduling found by the hybrid approach has helped to obtain the minimum production cost compared to the other monotonous methods. This proves the main feature of the hybridization technique [32] that allows the combination of the advantages of various methods. Indeed, the tabu search, as table 4 shows, is found the most efficient regarding the convergence time but it has the highest total production cost. As for the Lagrangian relaxation, has allowed to have a very high production cost and requires a considerable time to converge. Whereas, the particle swarm optimization is conducted in a good convergence time but the cost is high compared to our TS-PSO-LR strategy. This comparison reflects the performance of the hybrid strategy, both in production cost and in convergence time.

Therefore, based on Figure 3, we can notice that the total amount of generated power by the production units is very similar to that consumed with a very limited amount of spinning reserve power compared to PSO, TS and LR methods where the generated powers are much higher than the amount requested. This proves the effectiveness of tracking of the consumed power per each hour and shows the performance of the algorithms enabling to get a minimal production cost. Besides, this minimal production cost has been established thanks to a good On/Off statements scheduling set for each production units (Figures 4 and 5). The organization is made through an estimation of the amount of load desired by the electric network, while taking into account of the allowable constraints.

We note that the optimal unit commitment scheduling found by our PSO-TS-LR strategy is characterized by the On status of the powerful unit (615 MVA) [3,6,8,9,13] throughout 48 hours and this is due to the minimum up and the minimum down-time constraints and to the power demand governed for each hour. We note that the unit commitment scheduling of the second production unit (370 MVA) is respecting the during minimum up/down time equal to 8 hours which proves the effectiveness of the control

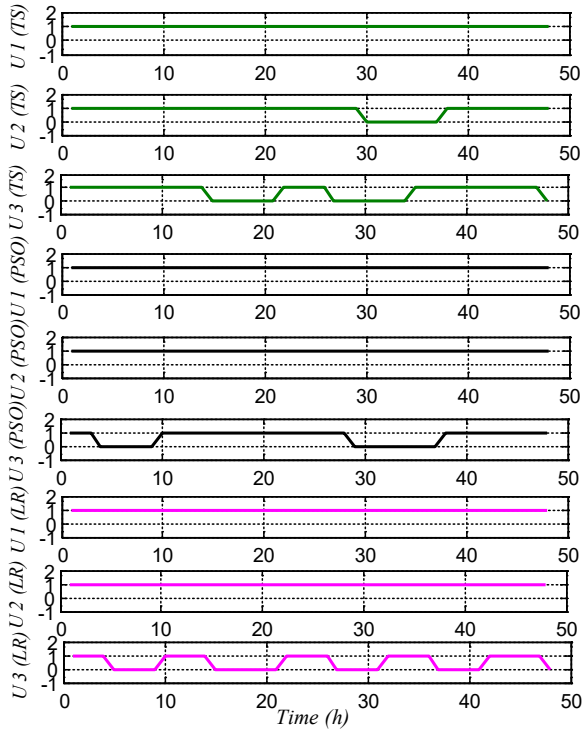


Figure 4. Optimal binary combination of units operation through TS method, PSO method and LR method

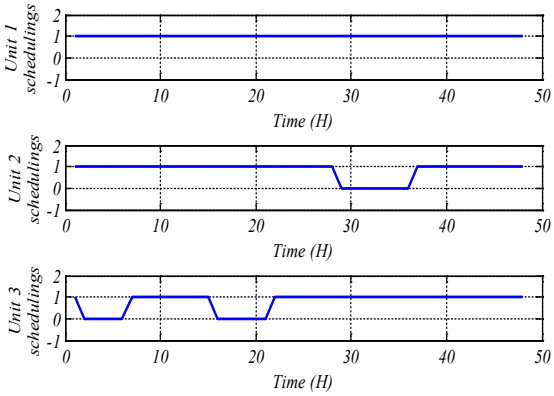


Figure 5. Optimal binary combination of units operation through hybrid TS-PSO-LR method

strategy. The third production unit undergoes Off status according to the production requirement and according to the power demand.

Furthermore, we confirm that our approach has allowed to select precisely the production units that should be available to respond to the load demand of the electrical network over a future period.

In addition, the adopted approach was promising both in terms of convergence to get the best optimal solutions to minimize the production cost and for an efficient unit commitment scheduling for the different production units, figure 6. Our strategy differs from other evolutionary computing techniques in providing an acceptable solution within a relatively short time and is likely to lead the search towards the most promising solution area.

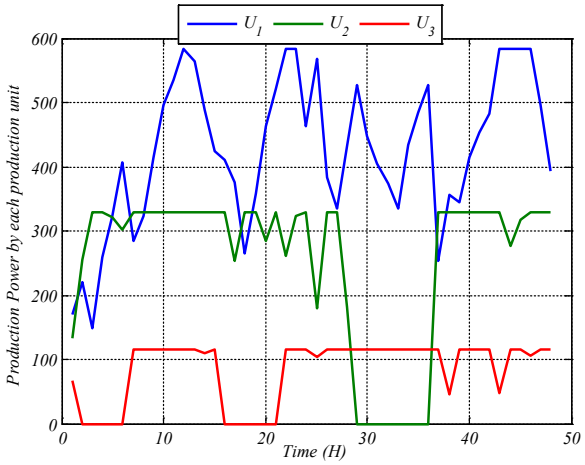


Figure 6. Generated Power by each production unit during 48 hours

6. Conclusion

This work shows the implementation of a new hybrid strategy which combines between Tabu Search, Particle Swarm Optimization and Lagrangian Relaxation. Herein, we have benefited from the rapidity of Lagrangian relaxation method, the storage solution with the memory of Tabu Search method and the flexibility to find optimal solution given by the Particle Swarm Optimization method. The proposed strategy TS-PSO-LR has presented high performances in optimizing the production cost and a capability of convergence to a global optimum as quick as possible compared to meta-heuristic (PSO, TS) and deterministic (LR) methods. In addition, the proposed strategy has ensured a proper unit commitment scheduling leading to get a minimal production cost. The right choice of the initial population suggests the possibility to obtain improvements in execution time. In addition, our strategy provides a fast enough time to converge to the optimal solution; which demonstrates the effectiveness of the adopted strategy compared to that obtained by Lagrangian Relaxation method and Particle swarm optimization methods.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Merlin, A., Sandrin, P. (1983) "A new method for unit commitment at Electricity De France", IEEE Trans on Power Apparatus and Systems; 102:1218-25.

[2] Lin, FT., Kao, C., Hsu, C. (1993) "Applying the genetic approach to simulated annealing in solving some NP-Hard problems". IEEE Trans on Systems, Man and Cybernetics; 23(6):1752-67.

[3] Sahbi Marrouchi, Moez ben hessine and Souad Chebbi, "New strategy based on Combined Use of Particle Swarm Optimization and Gradient methods to solve the Unit Commitment Problem", 15th IEEE International Conference on Environment and Electrical Engineering (EEEIC), Rome, Italy, 10-13 Juin 2015.

[4] Chunlin, D., Liu, Y., (2014) "Sample average approximation method for the chance-constrained stochastic programming in the transportation model of emergency management", Int. J. of Simulation and Process Modelling, Vol.9, No.4, pp.222 – 227.

[5] Richard, L. W., David, H. C. (2015) "A comprehensive method for solving finite-state semi Markov processes", Int. J. of Simulation and Process Modelling, Vol.10, No.1, pp.89 – 99.

[6] Rabeh, A., Sahbi, M., Moez, B.H., Houda, J., Souad, C., (2012) "Voltage Control Strategy of an Electrical Network by the Integration of the UPFC Compensator", International Review On Modelling and Simulation, vol. 5, no. 1, pp. 380–384.

[7] Galina, M., Vitaly, B., (2014) "Integrated planning and scheduling built on cluster analysis and simulation optimization", Int. J. of Simulation and Process Modelling, Vol.9, No.1/2, pp.81 – 91.

[8] Moez, B.H., Souad, B.S, (2014) "Accurate Fault Classifier and Locator for EHV Transmission Lines based on Artificial Neural Networks", Mathematical Problems in Engineering (MPE), ID 240565, vol 2014.

[9] Sahbi, M., Souad, C., (2014) "Combined Use of Particle Swarm Optimization and Genetic Algorithm Methods to Solve the Unit Commitment Problem", 16th International conference on Sciences and Techniques of Automatic control & computer engineering (STA), Monastir, Tunisia.

[10] Rahul, G., and Sharma, A.K. (2008) "Economic generation and scheduling of power by genetic algorithm", Journal of Theoretical and Applied Information Technology (JATIT).

[11] Aoki, K., Satoh, T., Itoh, M., Ichimori, T., and Masegi, K. (1987) "unit commitment in a large-scale power system including fuel constrained thermal and pumped-storage hydro", IEEE Transactions on Power Systems, Vol. PWRS-2, No. 4.

[12] Ouyang, Z., Shahidehpour, S. M. (1991) "An intelligent dynamic programming for unit commitment application", IEEE Trans on Power Systems; 6(3):1203-9.

[13] Sahbi, M., Souad, C., (2014) "A Comparative Study of Fuzzy Logic, Genetic Algorithm, and Gradient-Genetic Algorithm Optimization Methods for Solving the Unit Commitment Problem", Mathematical Problems in Engineering, vol. 2014, Article ID 708275, 14 pages, 2014. doi:10.1155/2014/708275

[14] Sasaki, H., Watanabe, M., Kubokawa, Yorino, J. N. and Yokoyama, R. (1992) "A solution method of unit commitment by artificial neural network", IEEE Transactions on Power Systems, Vol. 7, No. 3.

[15] Sahbi, M., Souad, C., (2014) "Unit Commitment Optimization Using Gradient-Genetic Algorithm and Fuzzy Logic Approaches", Complex system modeling and control through intelligent soft computations, Vol. 319, Springer book.

[16] Mantawy, A. H., Youssef, Y. L., Magid, L. A., and Shokri, S. Z., Selim, Z. (1998) "A unit commitment by Tabu search," Proc. Inst. Elect. Eng.Gen. Transm.Dist., vol. 145, no. 1, pp. 56–64.

[17] A. H. Mantawy. L. and Abdel MagidS. Z. Selim, (1998) "A unit commitment by Tabu search'', 1998 IEEE Proceedings online no. 1998.

[18] C. P. Cheng, C. W. Liu, C. C. Liu. (2002) "Unit commitment by annealing-genetic algorithm, Electrical Power and Energy Systems, vol. 24 pp 149-158.

[19] Logenthiran, T. and Dipti, S., (2010) ''Particle Swarm Optimization for Unit Commitment Problem'', International Conference on Probabilistic Methods Applied to Power Systems, pp 642 - 647.

[20] Charles, G., Christober A. R., (2009) ''Neural Based Tabu Search Method for Solving Unit Commitment Problem with Cooling-Banking Constraints'',Serbian Journal of electrical engineering Vol. 6, No. 1, May 2009, 57-74 UDK: 004.832.2, 2009.

[21] Senthil, K. and Palanisamy, V., (2006) ''A New Dynamic Programming Based Hopfield Neural Network to Unit Commitment and Economic Dispatch'' International Conference on Industrial Technology,pp 887 – 892, Mumbai.

[22] Chuan-Ping Cheng, Chih-Wen Liu, and Chun-Chang Liu, ''Unit Commitment by Lagrangian Relaxation and Genetic Algorithms'', IEEE Transactions on Power Systems, vol. 15, NO. 2, MAY 2000.

[23] C. Christober Asir Rajan and M. R. Mohan,"An Evolutionary Programming-Based Tabu Search Method For Solving The Unit Commitment Problem," IEEE Transactions on Power Systems, VOL. 19, NO. 1, FEBRUARY 2004.

[24] S. Senthil Kumar and V. Palanisamy, ''A New Dynamic Programming Based Hopfield Neural Network to Unit Commitment and Economic Dispatch'',IEEE Transactions on Power Systems, 2006.

[25] Charles Gnanakkan Christober Asir Rajan, ''Neural Based Tabu Search Method for Solving Unit Commitment Problem with Cooling-Banking Constraints'', Serbian Journal of electrical engineering Vol. 6, No. 1, May 2009, 57-74 UDK: 004.832.2, 2009.

[26] Christober C. Asir Rajan, ''Genetic algorithm based tabu search method for solving unit commitment problem with cooling — banking constraints'', Journal of electrical engineering, vol. 60, NO. 2, 2009.

[27] Alma Ademovic, Smajo Bisanovic and Mensur Hajro, '' A Genetic Algorithm Solution to the Unit Commitment Problem Based on Real-Coded Chromosomes and Fuzzy Optimization'', IEEE Transactions on Power Systems, 2010.

[28] Zhao, B., Guo, C. X., Bai, B. R., and Cao, Y.J. (2006) " An improved particle swarm optimization algorithm for unit commitment", International Journal of Electrical Power and Energy Systems, vol. 28, no. 7, pp. 482–490.

[29] Raglend, I.J., Raghuveer, C., Avinash, G.R., Padhy, N.P., and Kothari, D. P. (2010) "Solution to profit based unit commitment problem using particle swarm optimization", Applied Soft Computing Journal, vol. 10, no. 4, pp. 1247–1256.

[30] Chuan,-P.C., Chih, W.L., and Chun, C.L., (2000) ''Unit Commitment by Lagrangian Relaxation and Genetic Algorithms'',IEEE Transactions on Power Systems, vol. 15, NO. 2.

[31] Sahbi, M., Souad, C., (2014) "New strategy based on Fuzzy-Logic Approach to Solve the Unit-Commitment Problem", International Conference on Control, Engineering & Information Technology (CEIT), Monastir, Tunisia.

[32] Letizia, N.; Alessandro, C., Carlos, A., Rafael, D., (2014) "Hybrid approach for container terminals performances evaluation and analysis", Int. J. of Simulation and Process Modelling, Vol.9, No.1/2, pp.104–112.

.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 366-372 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Non-rigid Registration for 3D Active Shape Liver Modeling

Nesrine Trabelsi[*,1], Mohamed Ali Cherni[2], Dorra Ben Sellem[1,3,4]

[1] *Université de Tunis El Manar, Institute of Medical Technologies of Tunis , LR13 ES07 BTM, Tunis 1060 Tunisia*

[2] *Université de Tunis, LR13 ES03 SIME, ENSIT, Montfleury 1008 Tunisia*

[3] *Université de Tunis El Manar, Faculté de Médecine de Tunis, Tunis 1007 Tunisia*

[4] *Institut Salah AZAIEZ, Service de Médecine Nucléaire, Tunis 1006 Tunisia*

ARTICLEINFO

Article history:
Received: 10 October, 2017
Accepted: 18 January, 2018
Online: 10 February, 2018

Keywords:
3D Active Shape Model
B-Spline registration
CT-liver-scan
iso-surfaces
Marching Cubes
ROC-curves analysis

ABSTRACT

To avoid biopsies, doctors use non invasive medical techniques such as the computed tomography. Even that, the detection of the liver remains a big challenge because of the gray level and shape variations which depend on patients and acquisition modalites. In this work, we propose to create a 3D liver model in the training phase of 3D active shape model algorithm. This training model will be deformed according to any given 3D data for liver segmentation. The contribution of our work is the use of the Non-rigid registration with a B-spline registration on the training phase. We tested our method on an open access database ("3D-IRCADb") and on our database obtained from the radiology department of the National Oncology Institute of Tunis. Both data-sets showed the reliability of the method with an accuracy equal to 69.98% and 71.18% respectively for our database and "3D-IRCADb".

1 Introduction

Computed Tomography (CT) is a non invasive technique with a large field of view. It helps doctors to detect some hepatic disease. First, they identify the liver from the abdominal CT slice using two phases of enhancement: portal and arterial. In each time of enhancement, the gray level of the liver changes. In this work, we propose a 3D method to extract the liver from the CT slice according to its shape. We use different 3D CT exam with different phases of enhancement containing normal and pathological cases. The aim of our work is to use an advanced technique of segmentation to extract liver from images of any type of modality. For that reason, we propose to use the Active Shape Model (ASM). The ASM have been recognized as robust solution for a supervised technique of segmentation. Our contribution is to improve the performance of the 3D ASM algorithm with a pre-processing phase using a non rigid registration based on B-Spline transformation. The following paper is an extension of work originally presented in the 7th International Conference on Sciences of Electronics, Technologies of Information Telecommunication" [1]. In this work, we improve the evaluation of results using the ROC-Curve analysis and we ensure the reliability of our method using our database in addition to the "3D-IRCADb". Also, we compare the results not only with the Isosurfaces also with the Marching Cubes 3D reconstruction. The paper is organized as follow: Section 2 presents some literature related works. In section 3, we explain the 3D-ASM method and its different steps for 3D-model construction. the proposed method is applied on the open access database "3D-IRCADb" and our database. In section 4, we use the ROC-Curve analysis to evaluate the 3D liver modeling. First, we compare our 3D model with the Isosurfaces 3D reconstruction, then, with the Marching Cubes technique. Finally, we resume our proposed method and the obtained results on the conclusion.

2 Related works

The major point of interest of the Computed-aided diagnostic system is to solve difficulties of the medical field. The segmentation technique is the most important step in those system. It allows the physicians to extract the region of interest and to analyze it for more helpful medical information. In this con-

[*]Corresponding Author: Nesrine Trabelsi, nesrine.trabelsi@istmt.utm.tn

https://dx.doi.org/10.25046/aj030145

text, many techniques of segmentation have been developed such as morphological operations exposed on the work of Pham The Bao et al.[2]to extract the liver from the volumetric CT images. In [3], they improved morphological operator algorithm by using the watershed technique to segment the micro-tomographic trabecular bone. The watershed algorithm is widely used to identify the region of interest. In [4],P. Rodrigues et al. used it with an open software "Mevislab" to extract the liver from the CT exam with a 87% of accuracy. In [5], the author used the Fisher algorithm to improve the watershed algorithm for the radar images segmentation. Moreover, there are some intensity based techniques which detect the object according to the intensity repartition. In [6], the author constructed a model of the intensity distribution for the surface of the liver, cysts and lesions. Then, they calculated the probability of pixel belonging to each classification of different regions. In [7], Alom et al. used the "SGM growing slice method", an algorithm for 3D segmentation of the liver. Also, we note that algorithms of classification are used to segment the liver. For example, in the work [8], authors made a new texture feature extraction to improve liver classification using K-mean algorithm. While, in [9], the author used the "Support Vector Machine" algorithm and the surface distance maps for 3D-liver segmentation from CT-scans. Previously exposed works are based on the intensity of the liver and its gray level. In order to improve the segmentation technique, we propose a 3D-supervised-method for liver segmentation to overcome problems of gray level variation.

3 Methods and materials

In this section, we will explain the use of the shape context based on non-rigid surface registration with a B-Spline transformation in order to adapt the Active Shape Model for 3D CT liver segmentation. We start by a global description of the proposed method. Then, we detail each step. Figure 1 describes the different steps of our work to segment 3D CT liver. First, we note that the cases of the two databases are acquired with different parameters. So each exam has different number of slices, resolution and voxel size. To create the 3D model of the liver using the 3D ASM, we must uniform the volumetric data in the pre-processing phase. Several works combine different registration algorithm with a 3D segmentation method to extract a volumetric data such as to segment the left ventricle in [10] and the mandibular canal in [11]. In our work we propose to use a non-rigid registration with a B-Spline transformation. After the pre-processing phase, the 3D CT-scan has the same size of the matrix of vertices and the matrix of faces. Those two matrix will be used on the training phase of the 3D-ASM. Finally, we create the 3D liver model in the testing phase of the 3D-ASM.

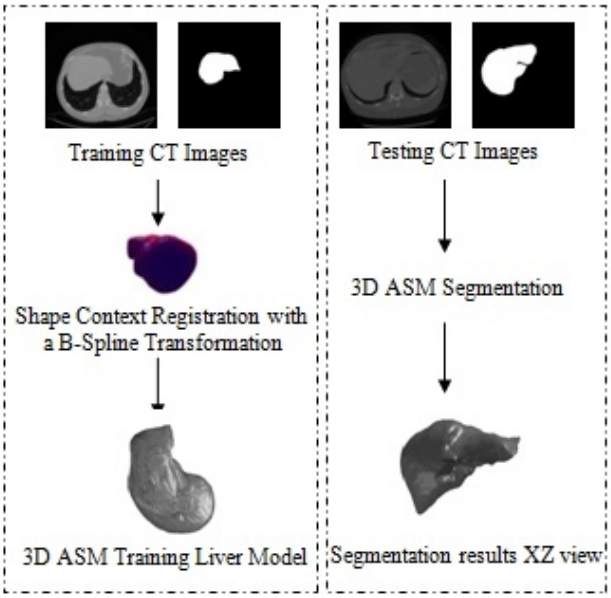


Figure 1: Diagram of the main work for 3D-ASM Liver segmentation

3.1 3D pre-processing CT-data

The CT-Scan measure the attenuation of the X-ray crossing the organ. These measurements are taken from different angles to produce a cross-sectional image called also slice separate with a thickness value. The total number of slices makes a volume information of the explored organ. However, doing the CT-Scan exam allows doctor to see details of the scanned organ in multi-planes (Coronal, axial and transverse). The obtained CT image is a matrix composed in pixels. Considering the the thickness between slices, each pixel represents a small volume element called also voxel. Its size depends on the matrix size, the thickness and the field of view. The CT images will be next stored as a numeric image. We note that the CT-images are coded according to thee Hounsfield units (UH) expressed by equation 1:

$$HU = \frac{\mu_{tissu} - \mu_{watter}}{\mu_{watter}} \quad (1)$$

In our work,we first change the UH to a gray level scaling using the DICOM header. We can define the gray value of a pixel as follow in equation (2):

$$Graylevel = \frac{UH \pm RescaleIntercept}{RescaleSlope} \quad (2)$$

where: The rescale intercept is equal to 0 for the IRCADb database and −1024 for our database; The rescale slope is equal to 1 for the both databases.
After changing the UH to the gray level scale, we uniform the slice number of the training data so all the training exams will have the same number of slices. We add a black slice at the end of the CT exam to reach the desired number. Then, we start the pre-processing step to achieve the same matrix size of

faces and vertices to all the training 3D data using the Shape Context Correspondence Point Model algorithm. It is based on a registration method. We can classify the registration into two classes: rigid and non-rigid transformation [12–14]. In our work, we employ a diffeomorphic B-Spline cubic transformation for no-rigid registration. We note : "C_1" a contour of an object 1 and "p_i" the number of points selected to define the contour; "C_2" a contour of an object 2 and "q_j" the number of points selected to define this contour. Then, we measure the cost function "C"between the two objects. It is defined in (3) as follow:

$$C(p_i, q_j) = \frac{1}{2} \sum_{k=1}^{k} \frac{[h_i(k) - h_j(k)]^2}{[h_i(k) + h_j(k)]} \quad (3)$$

With: h_i and h_j respectively histograms of the two shapes. After calculating the cost function, we minimize it to increase the degree of the similarity and the matching between the two shapes. To minimize the cost function, we choose a type of transformation in the registration phase. The type of transformation depends on the feature to be extracted from the object [15, 16], such as the match point algorithm[17], the 3D coordinate[18]and the intensity transformation using an iterative closet point algorithm [19, 20]. Figure 2 shows the different steps of the pre-processing phase. The first step of the registration phase using the Shape Context algorithm is to define the type of the transformation. In this work, we use the B-Spline cubic transformation. Then, we choose the mastery data which have the biggest number of slice and a floating data. This transformation is based on the surface matching with a spatial alignment of the two volumetric CT data. As an output of this phase, the CT exam will have the same size of vertices and faces matrix in order to be used on the next stage which is the 3D-ASM.

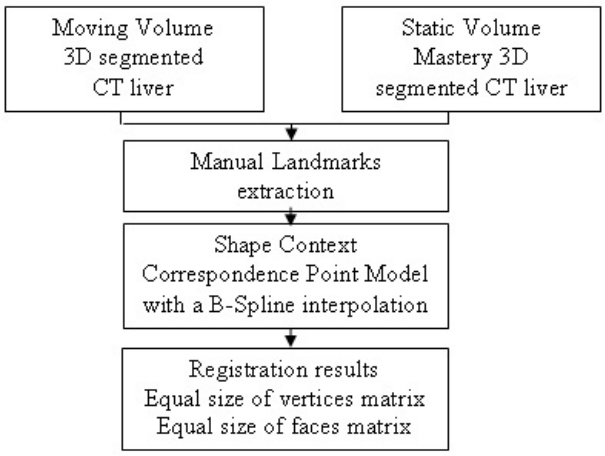


Figure 2: Pre-processing phase using Shape Correspondence Point Model

3.2 3D Active Shape Model

The 3D-ASM is an iterative supervised segmentation algorithm firstly proposed in [21] in 2002. It contains two phases: training and testing. Thanks to this algorithm based on the distribution point model (PDM) [22], we train our system to learn different liver shape variations. The input data in the training phase are 3D surface matrix containing the faces, the vertices of the 3D mesh and the binary volume of the extracted liver from the CT slice. At the beginning, we define for an object "O" a set of target point called landmark "n". The "n" points reform a vector of shape noted $X = (x, x_2, ..., x_n, y_1, y_2, ..., y_n, z_1, z_2, ..., z_n)^T$. After the identification of vector shape, we calculate from the 3D input data the covariance matrix. Then, we use the principal analyze components (PCA)[23]to identify the eigen-vectors"'x_i and their eigen-values. The PCA allows us to calculate the mean shape and its variations. Finally, we apply the PDM algorithm to create our training shape model noted $\tilde{X}$ in (4).

$$\tilde{X} = \bar{X} + \phi_s b_s \quad (4)$$

With: $\bar{X}$: aligned training shape; b_s:shape parameter vector of the shape. It is equal to $b_s = \pm m \sqrt{\lambda_i}$; $m \leq 3$; ϕ_s: eigenvectors corresponding to the eigenvalue of the covariance matrix. The next step in the 3D-ASM algorithm is to update the training model using the Mahalanobis distance[24]. Then, we apply the training deformed model to create our 3D liver model according to a given CT-data without doing a registration step.

4 Results and discussion

In this section, we start by a description of the databases used in this work and we present the obtained results. In order to evaluate the accuracy of our 3D liver modeling, we use the ROC-Curve analysis.

4.1 Description of databases

The proposed method was tested on two databases: our database (15 CT-exams) and the "3D-IRCADb" one (20 CT-exams). In [25], twenty radiologists expert describe the different parameters of the 3D IRCADb database. We found that the voxel size is various from one case to another from $0.57cm^3$ to $1.6cm^3$. Also in our database acquired at the radiology department of National oncology Institute "Salah AZAIEZ" of Tunis, the voxel size varies from 1.25cm*1.25cm*1.25cm to 1.4cm*1.4cm*1.4cm. For more efficiency, we use for the training phase, the 3D CT-scans with the same voxel size. In this work,for each database, we use the 80% of the CT-scan data having the same slice thickness for the training phase and 20% of 3D CT-scan for the testing phase.

4.2 Experiments

This study presents the use of the 3D-ASM for CT liver segmentation. For the beginning, we make a registration phase to recover some parameters such as rotation, scaling and translation in order to minimize the dissimilarity between two 3D CT-data. Figure 3 shows 3D mesh of the diffeomorphic B-Spline cubic registration. After the Shape Context registration, the output of this stage is used as input of the training phase of 3D-ASM algorithm. Different results of the 3D training shape are exposed in Figure 4.

Figure 3: 3D liver mesh of the B-Spline registration

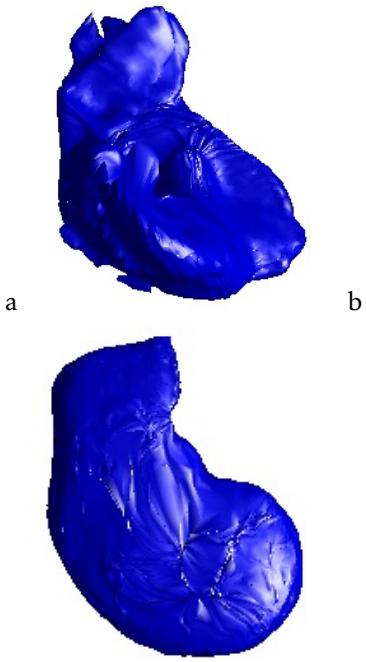


Figure 4: Training 3D model:a.Case from our database; b.Case from 3D-IRCADb database

To establish the proposed work, we use for the training phase of the 3D-ASM the cases which have the same thickness between slices and total number of slice. Then, we handle the corresponding Shape Context Points Model with a B-Spline transformation to create the same size of a 3D Liver grid. After the preprocessing step, we apply the 3D-ASM algorithm. To evaluate, in this work, the obtained results, we make a 3D reconstruction of the liver using the Isosurfaces and the Marching Cubes method.Then we compare them with the 3D-ASM of the liver. Figure 5 and figure 6 show different views of the 3D liver shape obtained from our database and the "3D-IRCADb" open access database.

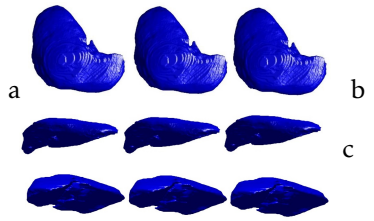


Figure 5: 3D liver segmentation of cases from "3D-IRCADb" database using respectively from the left to the right: the Isosurfaces, the Marching Cubes and the proposed method: a.XY view; b.XZ view; c.YZ view

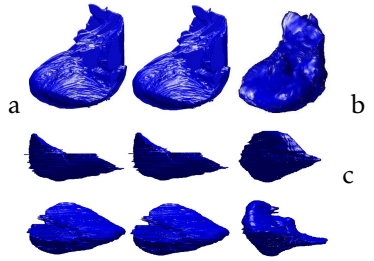


Figure 6: 3D liver segmentation of cases from our database using respectively from the left to the right: the isosurfaces, the Marching Cubes and the proposed method: a.XY view; b.XZ view; c.YZ view

After visual comparison, we use five parameters essentially used on ROC-curves analysis widely used in evaluation on the medical field [26]. These parameters are: sensitivity (SN), specificity (SP), accuracy (ACC), positive predicted value (PPV), negative predicted value (NPV) and area under the curve (AUC). We note the V_s is the segmented volume and V_{GT} the ground truth one. We define the True Positive(TP) as the degree of matching between the 3D mesh of V_s and the V_{GT} expressed in(5).

$$TP = V_s \cap V_{GT} \tag{5}$$

Its complement called False Negative (FN) is defined in (6).

$$FN = V_{GT} - TP \tag{6}$$

Moreover, we define the True Negative (TN) when we found a similarity on the false information. It is given in (7).

$$TN = 1 - (V_s \cup V_{GT}) \tag{7}$$

The False Positive (FP) is when our system detect a falsse mesh similiarity when there is no resemblance. It is given in (8).

$$FP = V_s - TP \tag{8}$$

According to those parameters, we calculate the sensitivity given in (9), the specificity given in (10) and the accuracy given in (11).

$$Sensitivity = \frac{TP}{TP+FN} \tag{9}$$

$$Specificity = \frac{TN}{TN+FP} \tag{10}$$

$$Accuraccy = \frac{TN+TP}{TN+TP+FN+FP} \tag{11}$$

We expose in tables 1 and 2 the different obtained values of the previously defined parameters. In this work, for each database, the liver was segmented by experts. So, we make a 3D reconstruction of these segmented volume using Isosurfaces and Marching Cubes techniques. Then, we apply the ROC-Curve independently. In the first hand, we apply it between our 3D results and 3D isosurfaces. On the second hand, we extract the different measure of the ROC-Curve between our 3D results and the Marching Cubes reconstruction.

	Our database	3D-IRCADb
Sensitivity (%)	68.38	63.8
Specificity (%)	71.57	59.36
PPV(%)	69.88	60.23
NPV (%)	70.85	59.52
Accuracy (%)	69.98	59.51
AUC	0.74	0.63

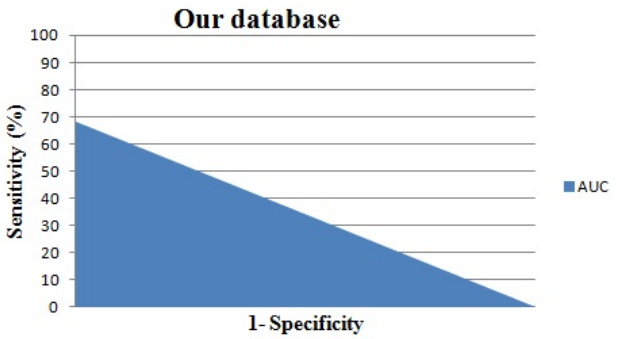
Table 1: Performances of the 3D-ASM with a B-spline registration versus the Isosurfaces reconstruction

	Our database	3D-IRCADb
Sensitivity (%)	75.16	50.06
Specificity (%)	65.93	49.94
PPV(%)	69.16	50.07
NPV (%)	72.33	50.76
Accuracy (%)	70.44	50.07
AUC	0.74	0.49

Table 2: Performances of the 3D-ASM with a B-spline registration versus the Marching Cubes reconstruction

Comparaing to the Isosurfaces and Marching Cubes methods, the 3D-ASM based on the Shape Context registration with a B-Spline registration achieve about 70% of accuracy and 71,57% of specificity. Those promising results, ensure the efficiency of the proposed method based on a non-rigid registration for the pre-processing phase. We found that we achieve better results with our database.

(a)



(b)

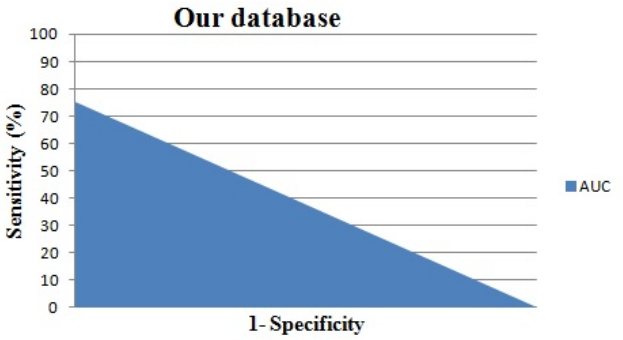
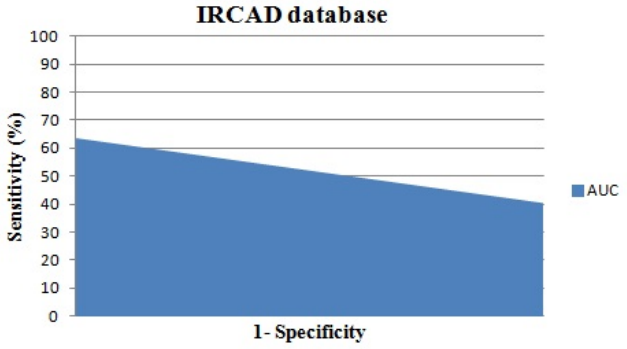


Figure 7: The area under the curve for cases from our database: a.versus Isosurfaces; b.versus Marching Cubes

(a)



(b)

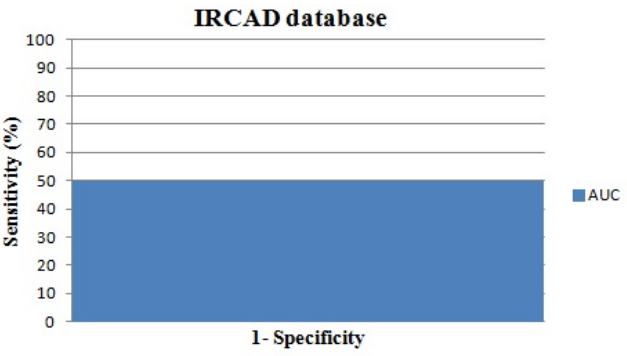


Figure 8: The area under the curve for cases from "3D-IRCADb"database: a.versus Isosurfaces; b.versus Marching Cubes

Our database contains pathological and normal cases. We used the normal cases for the training phase and we test the proposed method on pathological and normal cases. In fact, the 3D Active Shape Model is an advanced technique of 3D segmentation based on the shape. In this study, we prefer to use normals case with shape variation in order to create an efficient 3D learning model which will be deformed according to a given 3D CT-Scan exam. The inconvenient of the use of the pathological cases in the training phase is the no accurate liver shape since we can found one or several parts of the liver have been removed. The "3D-IRCADb" contains pathological cases without ablation which can be convincing for our study. To evaluate the efficiency of our proposed method using the "3D-IRCADb" database, we compare the 3D liver mesh obtained using our proposed method , the Isosurfaces and the Marching Cubes 3D reconstruction with the VTK files. The different results are exposed in the table 3. We found almost the same measure off the Area under the curve, presented in figure 9, but we achieve a slight improvement in terms of accuracy using the 3D Active Shape Model with a B-spline registration comparing to the Isosurfaces (Iso) and the Marching Cubes (MC).

	MC	Iso	Proposed Method
Sensitivity (%)	65.782	65.758	65.765
Specificity (%)	75.392	75.492	75.494
Accuracy (%)	70.629	70.648	71.182
PPV (%)	72.973	73.052	73.053
NPV (%)	68.775	68.744	68.736
AUC	0.757	0.756	0.751

Table 3: Comparison of the 3D liver mesh between the VTK files of "3D-IRCADb", the Isosurfaces, Marching Cubes and our proposed method using ASM with a B-spline registration.

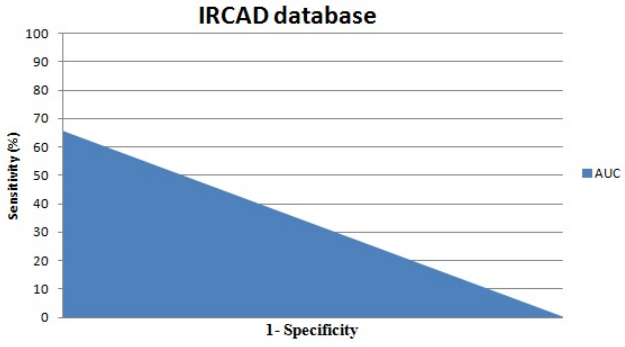


Figure 9: The Area under the curve: Comparison between our propose method and the VTK files of "3D-IRCADb" database

5 Conclusion

In this work, we used the Active shape Modeling for 3D CT-liver segmentation. In order to well create our model, we used in the pre-processing phase the Shape Context Corresponding Point Model with a B-Spline cubic interpolation. Evaluation of the proposed method has been performed by using 20 CT-exams from the "3D IRCADb"'database and 15 CT-exams from our database. We acheive a good results for 3D liver segmentation with a 70% of accuracy. Results obtained using our database are better than those obtained with the open access database "3D-IRCADb". This is due to the fact that the "3D-IRCADb" database is composed only by pathological cases contrariwise to our database which is composed by normal and pathological cases. As a conclusion, it would be better to use only normal cases or a much bigger number of normal cases than pathological cases to build a good training model. The proposed method brings to the doctors and the researchers, an approximate 3D model for normal liver cases. As a perspective, we attend to extract some feature from this 3D model for normal liver i order to detect liver pathologies according to the liver 3D mesh.

Conflict of Interest The authors declare no conflict of interest.

References

[1] N. Trabelsi, K. Aloui, and D. Ben Sellem. 3d active shape model for ct-scan liver segmentation. In *2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 161–165, Dec 2016. doi: 10.1109/SETIT.2016.7939859.

[2] Pham The Bao, Tran Hong Tai, Viet-Hang Duong, and Jia-Ching Wang. Liver segmentation from 3d abdominal ct images. In *Consumer Electronics-Taiwan (ICCE-TW), 2015 IEEE International Conference on*, pages 342–343. IEEE, 2015.

[3] W. A. Fourati and M. S. Bouhlel. Contour-based surface modeling and analysis of microtomographic trabecular bone images. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 423–427, March 2012. doi: 10.1109/SETIT.2012.6481951.

[4] P. Rodrigues, J. L. Vilaça, and J. Fonseca. An image processing application for liver tumour segmentation. In *1st Portuguese Biomedical Engineering Meeting*, pages 1–6, March 2011. doi: 10.1109/ENBENG.2011.6026097.

[5] B. Belkhaoui, A. Toumi, A. Khenchaf, A. Khalfallah, and M. S. Bouhlel. Segmentation of radar images using a combined watershed and fisher techniques. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 400–403, March 2012. doi: 10.1109/SETIT.2012.6481948.

[6] J. J. Chen, A. Kutluk, Y. T. Hu, and M. Hamit. Liver hydatid ct image segmentation based on localizing region active contours and modified parametric active contours. In *2014 7th International Conference on Biomedical Engineering and Informatics*, pages 217–221, Oct 2014. doi: 10.1109/BMEI.2014.7002773.

[7] M. Z. Alom, M. Mostakim, R. Biswas, and A. Chakrabarty. Automatic slice growing method based 3d reconstruction of liver with its vessels. In *Computer and Information Technology (ICCIT), 2013 16th International Conference on*, pages 338–344, March 2014. doi: 10.1109/ICCITechn.2014.6997361.

[8] J. Ma, F. Duan, and P. Guo. Improvement of texture image segmentation based on visual model. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 151–154, March 2012. doi: 10.1109/SETIT.2012.6481904.

[9] X. Zhang, J. Tian, D. Xiang, X. Li, and K. Deng. Interactive liver tumor segmentation from ct scans using support vector classification with watershed. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6005–6008, Aug 2011. doi: 10.1109/IEMBS.2011.6091484.

[10] C. Santiago, J. C. Nascimento, and J. S. Marques. Segmentation of the left ventricle in cardiac mri using a probabilistic data association active shape model. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7304–7307, Aug 2015. doi: 10.1109/EMBC.2015.7320078.

[11] F. Abdolali and R. A. Zoroofi. Mandibular canal segmentation using 3d active appearance models and shape context registration. In *2014 21th Iranian Conference on Biomedical Engineering (ICBME)*, pages 7–11, Nov 2014. doi: 10.1109/ICBME.2014.7043884.

[12] T. Huysmans, J. Sijbers, and V. Brigitte. Automatic construction of correspondences for tubular surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):636–651, April 2010. ISSN 0162-8828. doi: 10.1109/TPAMI.2009.93.

[13] M. Pereañez, K. Lekadir, I. Castro-Mateos, J. M. Pozo, Á. Lazáry, and A. F. Frangi. Accurate segmentation of vertebral bodies and processes using statistical shape decomposition and conditional models. *IEEE Transactions on Medical Imaging*, 34(8):1627–1639, Aug 2015. ISSN 0278-0062. doi: 10.1109/TMI.2015.2396774.

[14] Mohamed M Habib, Roshan A Welikala, Andreas Hoppe, Christopher G Owen, Alicja R Rudnicka, Adnan Tufail, Catherine Egan, and Sarah A Barman. Incorporating Spatial Information for Microaneurysm Detection in Retinal Images. *Advances in Science, Technology and Engineering Systems Journal*, 2(3):642–649, 2017. URL http://astesj.com/v02/i03/p82/.

[15] G. Pan, X. Zhang, Y. Wang, Z. Hu, X. Zheng, and Z. Wu. Establishing point correspondence of 3d faces via sparse facial deformable model. *IEEE Transactions on Image Processing*, 22(11):4170–4181, Nov 2013. ISSN 1057-7149. doi: 10.1109/TIP.2013.2271115.

[16] Olivier Commowick. *Design and Use of Anatomical Atlases for Radiotherapy*. Theses, Université Nice Sophia Antipolis, February 2007. URL https://tel.archives-ouvertes.fr/tel-00133432.

[17] M. Salehpour and A. Behrad. 3d face reconstruction by klt feature extraction and model consistency match refining and growing. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 297–302, March 2012. doi: 10.1109/SETIT.2012.6481932.

[18] A. Ghanbari, R. Abbasi-Asl, A. Ghaffari, and E. Fatemizadeh. Automatic b-spline image registration using histogram-based landmark extraction. In *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*, pages 1004–1008, Dec 2012. doi: 10.1109/IECBES.2012.6498119.

[19] Antonis D Savva, Theodore L Economopoulos, and George K Matsopoulos. Geometry-based vs. intensity-based medical image registration: a comparative study on 3d ct data. *Computers in biology and medicine*, 69:120–133, 2016.

[20] C. Xing and P. Qiu. Intensity-based image registration by nonparametric local smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2081–2092, Oct 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.26.

[21] B. van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever. Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, 21(8):924–933, Aug 2002. ISSN 0278-0062. doi: 10.1109/TMI.2002.803121.

[22] Timothy F Cootes, Christopher J Taylor, et al. Statistical models of appearance for computer vision, 2004.

[23] T. Tateyama, A. H. Foruzan, and Y. W. Chen. Pca based statical shape model of the spleen. In *2009 Fifth International Conference on Natural Computation*, volume 6, pages 36–39, Aug 2009. doi: 10.1109/ICNC.2009.695.

[24] Fadlalla G. Elfadaly, Paul H. Garthwaite, and John R. Crawford. On point estimation of the abnormality of a mahalanobis index. *Computational Statistics & Data Analysis*, 99:115 – 130, 2016. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2016.01.014. URL http://www.sciencedirect.com/science/article/pii/S0167947316000232.

[25] *Database IRCAD', Http://www.ircad.fr/fr/recherche/3d-ircadb-01-fr, Accessed 01 May 2015.* Http://www.ircad.fr/fr/recherche/3d-ircadb-01-fr, 2015.

[26] Enoch Opanin Gyamfi and Yaw Marfo Missah. Pixel-Based Unsupervised Classification Approach for Information Detection on Optical Markup Recognition Sheet. *Advances in Science, Technology and Engineering Systems Journal*, 2(4):121–132, 2017. doi: 10.25046/aj020417. URL http://astesj.com/v02/i04/p17/.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 373-383 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Adaptive observer design for a class of nonlinear systems with time delays

Ahlem Sassi[*1,2], Michel Zasadzinksi[1], Harouna Souley Ali[1], Kamel Abderrahim[2]

[1]*Centre de Recherche en Automatique de Nancy (CRAN, CNRS UMR 7039), Université de Lorraine, IUT de Longwy, 54400 Cosnes et Romain, France*

[2] *Commande Numérique des PRrocédés Industriels (CONPRI), National School of Engeneering in Gabes, Street Omar Ibn Elkhattab 6029 ZRIG Gabes, Tunisia*

ARTICLEINFO

Article history:
Received: 21 November, 2017
Accepted: 24 January, 2018
Online: 10 February, 2018

Keywords:
Time delays
Nonlinear system
Adaptive observer
LMI
Lipschitz function
Bilinear system

ABSTRACT

This paper deals with the design of adaptive observer for a class of nonlinear systems with time delays. Within this work, we develop an adaptive observer for a bilinear time delay system, then we extend those results in the presence of Lipschitz nonlinear functions in the system's dynamics. In the stability analysis of the estimation errors, we combine a Linear Parameter Varying (LPV) approach in the presence of time delays with a polytopic approach. The obtained stability conditions are given in terms of the solvability of Linear Matrix Inequalities (LMIs) on the vertices of a convex polytope. Numerical examples are finally given to show the effectiveness and feasibility of our results.

1 Introduction

This paper is an extension of work originally presented at the 6th International Conference on Systems and Control and entitled *"Full order adaptive observer design for time delay bilinear system"* [1]

During the last decades, several theoretical results with interesting applications were focused on the design of observers for nonlinear systems [2], [3], [4], [5], [6], [7] [8], [9] (and references there in). Due to the difficulty of setting the nonlinear behaviour in a system's dynamics and the non availability of all the state vector component, the design of observer is a challenging and open problem.

Indeed, since the presence of time delay is often encountered in the industrial processes, it is necessary to integrate it in the system model for a best modelization of these systems. The presence of time delay should not be neglected, because it can affect the systems performances and may lead, in certain cases, to its instability [10]. For these reasons, this class of systems has been intensively studied in the literature [11], [12], [13], [14] and will be considered also in the present work.

Apart the presence of time delays in a nonlinear model, another difficulty may appear: the presence of some unknown parameters, which should be taken into account in order to make the model more accurate. Hence, we propose in this work to design an adaptive observer, which does not only estimate the state vector of a nonlinear time delay system, but also the unknown parameters which affect its dynamics. Thanks to its ability to handle some challenging applications such as in robust and fault tolerant control, the adaptive observer allows to cope with the lack of knowledge on the system's unknown parameters. Classically, an adaptive observer may provide a suitable estimation of the states and the unknown parameters under some appropriate excitation conditions [3]. There are two major approaches to design adaptive observer. The first approach is based on the elaboration of an adaptation law derived from the stability analysis of a state observer. The convergence of the unknown parameters can be ensured under some persistent excitation [15], [16], [17]. The second approach consists in designing a state observer for an augmented system, where the state dynamics model is augmented with the dynamics of its unknown pa-

[*]Corresponding Author: Ahlem Sassi, CRAN, IUT de Longwy, 186 Rue de Lorraine 54400 Cosnes et Romain, France. Email:ahlemsessi@yahoo.com

https://dx.doi.org/10.25046/aj030146

rameters. [18], [19], [20], [21].

In this paper, we consider an adaptive observer design for a class of nonlinear time delay systems where the dynamics are nonlinear in the states and in the unknown parameters and contains some bilinear terms. The nonlinearities considered satisfied a Lipschitz condition. Here, the convergence of the adaptive observers is treated in a more tractable way and does not require to satisfy the persistent excitation condition as in [22], [23], [24] and contrary to many previous contributions. In a first time, an adaptive observer will be proposed in the absence of some nonlinearities. The considered system represents a time delay bilinear system affected by unknown parameters. The bilinear systems do appropriately model some physical processes than linear or nonlinear ones. Nevertheless, the fact of considering a bilinear model does not suffice, due to the presence of nonlinear behaviour which ought to be integrated. For that reason, we consider in a second time the presence of some additional Lipchitz nonlinear functions. It is necessary to study those results separately because the design of adaptive observer in the first case does not derive from the second one due to the presence of some unmeasured state variables in the nonlinear functions, which make the problem more conservative.

This paper is structured so that the statement of the problem and some useful formulas are presented in the section 2. Section 3 is devoted to the design of adaptive observer for a bilinear time delay system. This result is extended in section 4, by the addition of nonlinear Lipschitz functions in the delayed bilinear system dynamics. In section 5, a discussion of the obtained results is made by comparison with some previous ones. Finally, in section 6, the obtained results will be applied to numerical examples to show their effectiveness.

Notations. Throughout this paper, $\mathbb{R}^n$ denotes an n-dimentional Euclidean space, and $\|\cdot\|$ the associated Euclidean norm [25], where

$$\|x\| = \sqrt{x^T x}, \quad \forall x \in \mathbb{R}^n \tag{1}$$

and using expression (1), the following induced matrix norm given by

$$\|A\| = \max_{\|x\|\neq 0} \frac{\|Ax\|}{\|x\|} \tag{2}$$

is used in this paper where $A \in \mathbb{R}^{n\times m}$.
$(*)$ will denotes the transpose of the off-diagonal parts of a matrix.

2 Problem statement

In this work, a class of nonlinear time delay systems is investigated where the state space model is given by the following equation

$$\dot{x}(t) = A_0 x(t) + \sum_{i=1}^{m} A_i u_i x(t) + \ell(x,u) + A_{d_0} x(t-\tau_0) + \sum_{i=1}^{m} A_{d_i} u_i(t) x(t-\tau_i) + Bu(t) + Gg(x,u)\theta \tag{3a}$$

$$\dot{\theta} = 0 \tag{3b}$$

$$y = Cx(t) \tag{3c}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $\theta \in \mathbb{R}^q$, $y \in \mathbb{R}^p$ are the states vector, the input control, the unknown parameters vector and the output vector, respectively. τ_i for $i = 0,\dots,m$ are known constants delays. The matrices: $A_i \in \mathbb{R}^{n\times n}$, $A_{d_i} \in \mathbb{R}^{n\times n}$, for $i = 0,\dots,m$, $B \in \mathbb{R}^{n\times m}$, $G \in \mathbb{R}^{n\times r}$ and $C \in \mathbb{R}^{p\times n}$ are known with constant values.
The following assumptions are given and will be used in the sequel

Assumption 1. *The input $u(t)$ is bounded such that $u(t) \in \mathcal{U} \subset \mathbb{R}^m$, where*

$$\mathcal{U} = \{u : t \to \mathbb{R}^m / \forall t \in \mathbb{R}^+, u_{i,min} \leq u_i(t) \leq u_{i,max}, \mu_{i,min} \leq \dot{u}_i(t) \leq \mu_{i,max}\} \tag{4}$$

Assumption 2. *The function $\ell(x,u) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is Lipschitz in x, $\forall u \in \mathcal{U}$, i.e there exists a positive scalar b_1 such that*

$$\|\ell(x,u) - \ell(x^*,u)\| \leq b_1 \|x - x^*\| \tag{5}$$

The function $g(x,u) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{r\times q}$ is bounded $\forall x \in \mathbb{R}^n$ and $\forall u \in \mathcal{U}$, that is there exists a scalar b_g such that

$$\|g(x,u)\| \leq b_g$$

and is Lipschitz in x, $\forall u \in \mathcal{U}$, i.e there exists a positive scalar b_2 where

$$\|g(x,u) - g(x^*,u)\| \leq b_2 \|x - x^*\| \tag{6}$$

Assumption 3. *The unknown parameters are supposed to be bounded such that there exists a scalar $b_3 > 0$ verifying*

$$\|\theta\| \leq b_3$$

Since assumption 1 holds, we have put the inputs $u_i(t)$, for $i = 1,\dots,m$, and their derivatives in the same vector

$$\delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_m \\ \delta_{m+1} \\ \vdots \\ \delta_{2m}^T \end{bmatrix} = \begin{bmatrix} u_1 \\ \vdots \\ u_m \\ \dot{u}_1 \\ \vdots \\ \dot{u}_m \end{bmatrix} \tag{7}$$

one can see that the vector δ belongs to a convex polytope, described by

$$\mathcal{P} = [u_{1,\min}, u_{1,\max}] \times \dots \times [u_{m,\min}, u_{m,\max}] \times [\mu_{1,\min}, \mu_{1,\max}] \times \dots \times [\mu_{1,\min}, \mu_{1,\max}] \tag{8}$$

Let us note Φ the set of vertices of the convex polytope $\mathcal{P}$, where its cardinality is equal to 2^{2m}, and described by

$$\Phi = \{\sigma = [\phi_1,\dots,\phi_{2m}]^T \in \mathbb{R}^{2m} / \forall i \in [0,m], \phi_i \in \{u_{i,\min}, u_{i,\max}\} \text{ and } \forall i \in [m+1,2m], \phi_i \in \{\mu_{i,\min}, \mu_{i,\max}\}\} \tag{9}$$

In what follows, we point out the problem of the design of adaptive observer, which allows a simultaneous estimation of the states and the unknown parameters. As these observers are usually designed to be applied in robust and fault tolerant control, the add of the term $Gg(x,u)\theta$ in the system dynamics allows to model uncertainties affecting the system in the case of robust control, or may be used to model faults in the case of fault detection and isolation [18], [26], this term is more general than only the term $G\theta$ as considered in [1] and in the section 3 of the present work.

In a first time, we propose an adaptive observer for system (3) without considering the presence of the nonlinear Lipschitz functions, i.e $\ell(x,u) = 0$ and $g(x,u) = 1$. In a second time, a more general adaptive observer is proposed for the considered class of nonlinear time delay systems (3). The obtained results do not lead to the results of the section before (see the discussion in section 5), which justify the structure of this work.

Before starting the observer design, let us give some useful relations used in this paper. For $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, the following well known inequalities hold

$$2\|x\|\,\|y\| \leq \beta x x^T + \frac{1}{\beta} y y^T, \quad \forall \beta > 0 \tag{10}$$

$$xz^T + zx^T \leq cxx^T + \frac{1}{c} zz^T, \quad \forall c > 0 \tag{11}$$

3 Observer Design without Lipschitz nonlinearities

In this section, we consider system (3) without Lipschitz nonlinear functions, i.e. $\ell(x,u) = 0$ and $g(x,u) = 1$. Thus, an adaptive observer is proposed under the following form

$$\dot{\hat{x}}(t) = A_0\hat{x}(t) + \sum_{i=1}^{m} A_i u_i \hat{x}(t) + Bu(t) + A_{d_0}\hat{x}(t-\tau_0) + \sum_{i=1}^{m} A_{d_i} u_i(t)\hat{x}(t-\tau_i) + G\hat{\theta}(t) + L_x(y(t) - C\hat{x}(t)) \tag{12a}$$

$$\dot{\hat{\theta}}(t) = L_\theta(y(t) - C\hat{x}(t)) \tag{12b}$$

where $\hat{x}(t) \in \mathbb{R}^n$ and $\hat{\theta} \in \mathbb{R}^q$ are the estimated states vector and the estimated unknown parameters vector, respectively. $L_x \in \mathbb{R}^{n\times p}$ and $L_\theta \in \mathbb{R}^{q\times p}$ are the observer's gains to be determined.

As a full order observer, we propose this structure of the observer, since it is the most commonly used in the literature, and had shown performance results. Within this observer, only two observer gains have to be determined, contrary to the full order observer structure as proposed in [23], which may be cumbersome to compute, due to the number of matrices to be computed.

The estimation error has the following dynamics

$$\dot{e}(t) = H(u)e(t) + \sum_{i=0}^{m} H_{d_i} e(t-\tau_i) \tag{13}$$

where

$$e(t) = \begin{bmatrix} e_x(t) \\ e_\theta(t) \end{bmatrix} = \begin{bmatrix} x(t) - \hat{x}(t) \\ \theta(t) - \hat{\theta}(t) \end{bmatrix}$$

$$H(u) = \begin{bmatrix} \sum_{i=1}^{m} A_i u_i - L_x C & G \\ -L_\theta C & 0 \end{bmatrix} \tag{14}$$

$$H_{d_i}(u) = \begin{bmatrix} A_{d_i} u_i & 0 \\ 0 & 0 \end{bmatrix} \tag{15}$$

Notice that we considered $u_0(t) = 1$ for reasons of simplification. Then, system (12) is an adaptive observer for the delayed considered system described by (3) with $\ell(x,u) = 0$ and $g(x,u) = 1$, if and only if the estimation error system described by (13) is asymptotically stable. The stability of the estimation error $e(t)$ and the computation of the observer's gains L_x and L_θ are ensured via the following theorem.

Theorem 1. *Assume that assumption 1 holds. System* (12) *represents an adaptive observer to system* (3) *(with* $\ell(x,u) = 0$ *and* $g(x,u) = 1$*), and the estimation errors system* (13) *is quadratically stable for* $\sigma^j \in \Phi$, $j = 1,\ldots,2^{2m}$, *if there exist matrices*

- $P(\sigma^j) \in \mathbb{R}^{(n+q)\times(n+q)}$ *where*

$$P(\sigma^j) = P^T(\sigma^j) = \sum_{i=0}^{m} \sigma_i^j \begin{bmatrix} P_{i1} & P_{i2} \\ P_{i2}^T & P_{i3} \end{bmatrix} > 0 \tag{16}$$

  $P_{i1} \in \mathbb{R}^{n\times n}$, $P_{i2} \in \mathbb{R}^{n\times q}$ *and* $P_{i2} \in \mathbb{R}^{q\times q}$, *for* $i = 0,\ldots,m$,

- $M \in \mathbb{R}^{(n+p)\times(n+p)}$, *given by*

$$M = \begin{bmatrix} M_{11} & S_2 M_{22} \\ S_1 M_{11} & M_{22} \end{bmatrix} \tag{17}$$

  where $M_{11} \in \mathbb{R}^{n\times n}$ *and* $M_{22} \in \mathbb{R}^{q\times q}$ *are nonsingular matrices.* $S_1 \in \mathbb{R}^{q\times n}$ *and* $S_2 \in \mathbb{R}^{n\times q}$ *are some tuning matrices.*

- $Y_1 \in \mathbb{R}^{n\times p}$ *and* $Y_2 \in \mathbb{R}^{q\times p}$

and a positive scalar γ, *such that the following LMIs holds*

$$\begin{bmatrix} \alpha^j_{(1,1)} & \alpha^j_{(1,2)} & \alpha^j_{(1,3)} & \alpha^j_{(1,4)} & \alpha^j_{(1,5)} \\ * & \alpha^j_{(2,2)} & \alpha^j_{(2,3)} & \alpha^j_{(2,4)} & \alpha^j_{(2,5)} \\ * & * & \alpha_{(3,3)} & \alpha_{(3,4)} & \alpha^j_{(1,5)} \\ * & * & * & \alpha_{(4,4)} & \alpha^j_{(2,5)} \\ * & * & * & * & \frac{-1}{\gamma} I_k \end{bmatrix} < 0 \tag{18}$$

where

$$\alpha^j_{(1,1)} = \sum_{i=1}^{m} \sigma^j_{m+i} P_{i1} + \sum_{i=0}^{m} M_{11} A_i \sigma_i^j + \sum_{i=0}^{m} A_i^T M_{11}^T \sigma_i^j - Y_1 C - C^T Y_1^T - S_2 Y_2 C - C^T Y_2^T S_2^T + \sum_{i=0}^{m} M_{11} A_{d_i} \sigma_i^j + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T \sigma_i^j \tag{19a}$$

$$\alpha^j_{(1,2)} = \sum_{i=1}^{m} \sigma^j_{m+i} P_{i2} + M_{11}G + \sum_{i=0}^{m} A_i^T \sigma_i^j M_{11}^T S_1^T + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T S_1^T \sigma_i^j - C^T Y_2^T - C^T Y_1^T S_1^T \quad (19b)$$

$$\alpha^j_{(2,2)} = \sum_{i=1}^{m} \sigma^j_{m+i} P_{i3} + S_1 M_{11} G + G^T M_{11}^T S_1^T \quad (19c)$$

$$\alpha^j_{(1,3)} = \sum_{i=0}^{m} \sigma_i^j P_{i1} + \sum_{i=0}^{m} A_i^T M_{11}^T \sigma_i^j + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T \sigma_i^j - M_{11} - C^T Y_1^T - C^T Y_2^T S_2^T \quad (19d)$$

$$\alpha^j_{(1,4)} = \sum_{i=0}^{m} \sigma_i^j P_{i2} - M_{12} + \sum_{i=0}^{m} A_i^T M_{11}^T S_1^T \sigma_i^j - C^T Y_1^T S_1^T - C^T Y_2^T + \sum_{i=1}^{m} A_{d_i}^T M_{11}^T S_1^T \sigma_i^j \quad (19e)$$

$$\alpha^j_{(2,3)} = \sum_{i=0}^{m} \sigma_i^j P_{i2}^T - S_1 M_{11} + G^T M_{11}^T \quad (19f)$$

$$\alpha^j_{(2,4)} = \sum_{i=0}^{m} \sigma_i^j P_{i3} - M_{22} + G^T M_{11}^T S_1^T \quad (19g)$$

$$\alpha_{(3,3)} = -M_{11} - M_{11}^T \quad (19h)$$

$$\alpha_{(4,4)} = -M_{22} - M_{22}^T \quad (19i)$$

$$\alpha_{(3,4)} = -S_2 M_{22} - M_{11}^T S_1^T \quad (19j)$$

$$\alpha^j_{(1,5)} = [M_{11}A_{d_0}\sigma_0^j, 0, M_{11}A_{d_1}\sigma_1^j, 0, \ldots, M_{11}A_{d_m}\sigma_m^j, 0] \quad (19k)$$

$$\alpha^j_{(2,5)} = [S_1 M_{11} A_{d_0}\sigma_0^j, 0, S_1 M_{11} A_{d_1}\sigma_1^j, 0, \ldots, S_1 M_{11} A_{d_m}\sigma_m^j, 0] \quad (19l)$$

The observer gains are expressed by

$$L_x = M_{11}^{-1} Y_1 \quad (20a)$$

$$L_\theta = M_{22}^{-1} Y_2 \quad (20b)$$

Proof. The proof of the theorem 1 will be developed into two steps:

- In a first step, we give the stability conditions using a Lyapunov Krasovskii approach for LPV time delay systems, which leads to the resolution of an inequality.
- In a second step, we compute the observers matrices and we transform the obtained inequality into LMIs, using a Polytopic approach.

First Step. Let us consider a Lyapunov Krasovskii function candidate with this form

$$V(e) = \begin{bmatrix} e \\ \dot{e} \end{bmatrix}^T F(u) E \begin{bmatrix} e \\ \dot{e} \end{bmatrix} + \frac{1}{\gamma} \sum_{i=0}^{m} \int_{-\tau_i}^{0} \int_{t}^{t+\beta} \dot{e}^T(s)\dot{e}(s) ds d\beta \quad (21)$$

where $F(u) = \begin{bmatrix} P(u) & M \\ 0 & M \end{bmatrix}$, with $P(u) \in \mathbb{R}^{(n+p)\times(n+p)}$, $P(u) = P(u)^T > 0$ and M is a matrix with a structure described by (17). $E = \begin{bmatrix} I_{(n+q)} & 0 \\ 0 & 0 \end{bmatrix}$ and γ is a positive scalar.

Let us note $\varepsilon(t) = \dot{e}(t)$, and we rewrite system (13) under a descriptor form as follows

$$\begin{bmatrix} I_{(n+q)} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{e}(t) \\ \dot{\varepsilon}(t) \end{bmatrix} = \begin{bmatrix} 0 & I \\ H(u) + \sum_{i=0}^{m} H_{d_i} & -I \end{bmatrix} \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix} + \sum_{i=0}^{m} \begin{bmatrix} 0 \\ H_{d_i} \end{bmatrix} \int_{t}^{t-\tau_i} \varepsilon(\kappa) d\kappa$$

Then, using the latter equation, we differentiate the Lyapunov function in t, which leads to

$$\dot{V}(t) = \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix}^T \mathbb{H} \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix} + \sum_{i=0}^{m} \beta_i(t) - \frac{1}{\gamma} \sum_{i=0}^{m} \int_{t}^{t-\tau_i} \dot{e}^T(s)\dot{e}(s) ds \quad (22)$$

where

$$\mathbb{H} = \begin{bmatrix} h_{(1,1)} & h_{(1,2)} \\ * & h_{(2,2)} \end{bmatrix} \quad (23)$$

with

$$h_{(1,1)} = \dot{P}(u) + MH(u) + H^T(u)M^T + \sum_{i=0}^{m} MH_{d_i} + \sum_{i=0}^{m} H_{d_i}^T M^T$$

$$h_{(1,2)} = P(u) - M + H^T(u)M^T + H_{d_i}^T M^T$$

$$h_{(2,2)} = -M - M^T$$

and

$$\beta_i(t) = \begin{bmatrix} e^T(t) & \dot{e}^T(t) \end{bmatrix} F(u) \begin{bmatrix} 0 \\ H_{d_i} \end{bmatrix} \int_{t}^{t-\tau_i} \dot{e}(s) ds + \int_{t}^{t-\tau_i} \dot{e}^T(s) ds \begin{bmatrix} 0 & H_{d_i}^T \end{bmatrix} F^T(u) \begin{bmatrix} e(t) \\ \dot{e}(t) \end{bmatrix} \quad (24)$$

Using inequality (11), we majorate β_i as follows

$$\beta_i(t) \le \gamma \begin{bmatrix} e^T(t) & \dot{e}^T(t) \end{bmatrix} F(u) \begin{bmatrix} 0 \\ H_{d_i}(u) \end{bmatrix} \begin{bmatrix} 0 \\ H_{d_i}(u) \end{bmatrix}^T F^T(u) \times \begin{bmatrix} e(t) \\ \dot{e}(t) \end{bmatrix} + \frac{1}{\gamma} \int_{t}^{t-\tau_i} \dot{e}^T(s)\dot{e}(s) ds \quad (25)$$

Replacing inequality (25) in the expression of the derivative of the Lyapunov function $V(e)$ described by (22), implies

$$\dot{V}(e) \le \begin{bmatrix} e(t) \\ \dot{e}(t) \end{bmatrix}^T \begin{bmatrix} \overline{h}_{(1,1)} & \overline{h}_{(1,2)} \\ * & \overline{h}_{(2,2)} \end{bmatrix} \begin{bmatrix} e(t) \\ \dot{e}(t) \end{bmatrix}$$

where

$$\overline{h}_{(1,1)} = \dot{P}(u) + MH(u) + H^T(u)M^T + \sum_{i=0}^{m} MH_{d_i} + \sum_{i=0}^{m} H_{d_i}^T M^T + \gamma \sum_{i=0}^{m} MH_{d_i} H_{d_i}^T M^T$$

$$\overline{h}_{(1,2)} = P(u) - M + H^T(u)M^T + H_{d_i}^T M^T + \gamma \sum_{i=0}^{m} MH_{d_i}H_{d_i}^T M^T$$

$$\overline{h}_{(2,2)} = -M - M^T + \gamma \sum_{i=0}^{m} MH_{d_i}H_{d_i}^T M^T$$

The negativity of the derivative of the Lyapunov-Krasovskii function is equivalent to the following inequality:

$$\begin{bmatrix} \overline{h}_{(1,1)} & \overline{h}_{(1,2)} \\ * & \overline{h}_{(2,2)} \end{bmatrix} < 0$$

Applying the Schur complement on the latter inequality allows to get the following inequality,

$$\begin{bmatrix} a_{(1,1)} & a_{(1,2)} & a_{(1,4)} \\ * & a_{(1,3)} & a_{(1,4)} \\ * & * & \frac{-1}{\gamma} I_k \end{bmatrix} < 0 \tag{26}$$

where

$$a_{(1,1)} = \dot{P}(u) + MH(u) + H^T(u)M^T + \sum_{i=0}^{m} MH_{d_i}(u) + \sum_{i=0}^{m} H_{d_i}^T(u)M^T$$

$$a_{(1,2)} = P(u) - M + H^T(u)M^T + \sum_{i=0}^{m} H_{d_i}^T M^T$$

$$a_{(1,3)} = -M - M^T$$

$$a_{(1,4)} = [MH_{d_0}(u), \ldots, MH_{d_m}(u)]$$

with $k = (m+1)(n+q)$.

Second step. Now, using the polytopic approach, we consider that assumption 1 holds and that the inputs and their derivatives belong to the convex polytope $\mathcal{P}$ defined by (8). Then, using the vector δ given by (7), we rewrite the system's matrices, the Lyapunov matrix and its derivative as follows

$$P(u) = P(\delta) = P_0 + \sum_{i=1}^{m} \delta_i P_i, \tag{27a}$$

$$\dot{P}(u) = \overline{P}(\delta) = \sum_{i=1}^{m} \delta_{m+i} P_i \tag{27b}$$

$$H(\delta) = \begin{bmatrix} \sum_{i=0}^{m} A_i \delta_i - L_x C & G \\ -L_\theta C & 0 \end{bmatrix} \tag{27c}$$

$$H_{d_i}(\delta) = \begin{bmatrix} A_{d_i}\delta_i & 0 \\ 0 & 0 \end{bmatrix} \tag{27d}$$

Then, we compute inequality (26) in the whole set of the vertices of the polytope Φ. By taking the matrices $P^j(\sigma_i)$ and M with the form (16) and (17) respectively, we obtain the results given in the proof, where the observer's gains are given by (20). □

Now, as discussed in the introductory section, we consider in the next section a more general class of systems with Lipschitz nonlinearities.

4 Observer Design with Lipschitz nonlinearities

In this section, the objective is to design an adaptive observer for system (3) with the following general structure in order to estimate simultaneously the states vector x and the unknown parameters θ. For that reason, the following adaptive observer is considered

$$\dot{\hat{x}}(t) = \left(A_0 + \sum_{i=1}^{m} A_i u_i\right)\hat{x}(t) + Bu(t) + \ell(\hat{x},u) + A_{d_0}\hat{x}(t-\tau_0) + \sum_{i=1}^{m} A_{d_i}u_i(t)\hat{x}(t-\tau_i) + Gg(\hat{x},u)\hat{\theta}(t) + L_x(y - C\hat{x})$$

$$\dot{\hat{\theta}} = L_\theta(y - C\hat{x}) \tag{28}$$

where $\hat{x} \in \mathbb{R}^n$ and $\hat{\theta} \in \mathbb{R}^q$ are the estimates of the states x and the unknown parameters θ respectively. $L_x \in \mathbb{R}^{n\times p}$ and $L_\theta \in \mathbb{R}^{q\times p}$ are the observer's gains.

The convergence analysis made in this section is different from the section above, due to the presence of the Lipschitz nonlinear functions $\ell(x,u)$ and $g(x,u)$.

In order to ensure the convergence of the proposed adaptive observer, we had to choose the gains L_x and L_θ which guarantee that the errors $e_x(t) = x(t) - \hat{x}(t)$ and $e_\theta = \theta - \hat{\theta}$, converge to zero, in other words, the estimated vectors $\hat{x}$ and $\hat{\theta}$ converge to their actual values. Thus, this section is devoted to obtain the stability conditions in term of LMI.

Let us give the dynamics of the estimation errors $e_x(t)$ and $e_\theta(t)$ as follows

$$\dot{e}_x(t) = \left(A_0 + \sum_{i=1}^{m} A_i u_i - L_x C\right)e_x(t) + (\ell(x,u) - \ell(\hat{x},u)) + A_{d_0}e_x(t-\tau_0) + \sum_{i=1}^{m} A_{d_i}u_i e_x(t-\tau_i) + G\left(g(x,u)\theta - g(\hat{x},u)\hat{\theta}\right) \tag{29a}$$

$$\dot{e}_\theta(t) = -L_\theta C e_x(t) \tag{29b}$$

After adding and subtracting the term $\rho G e_\theta(t)$ to equation (29a), where ρ is a positive scalar, we obtain the following augmented error system

$$\dot{e}(t) = H(u)e(t) + \sum_{i=0}^{m} H_{d_i}(u)e(t-\tau_i) + H_\ell + \overline{G}\,\overline{H}_g \tag{30}$$

where H_{d_i} is defined by (15) and

$$H(u) = \begin{bmatrix} \sum_{i=1}^{m} A_i u_i - L_x C & \rho G \\ -L_\theta C & 0 \end{bmatrix}$$

$$H_\ell = \begin{bmatrix} \ell(x,u) - \ell(\hat{x},u) \\ 0 \end{bmatrix}$$

$$\overline{H}_g = H_g - \begin{bmatrix} \rho e_\theta(t) \\ 0 \end{bmatrix}$$

$$H_g = \begin{bmatrix} g(x,u)\theta - g(\hat{x},u)\hat{\theta} \\ 0 \end{bmatrix}$$

$$\overline{G} = \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix}$$

with $u_0(t) = 1$.

The quadratic stability of the estimation error system is ensured via the following theorem

Theorem 2. *Assuming that assumptions 1, 2 and 3 hold. System* (28) *is an adaptive observer for system* (3)*, and the estimation error is quadratically stable for $\sigma^j \in \Phi$, if there exist matrices*

- $P(\sigma^j) \in \mathbb{R}^{(n+q)\times(n+q)}$ *where*

$$P(\sigma^j) = \sum_{i=0}^{m} \sigma_i^j \begin{bmatrix} P_{i1} & P_{i2} \\ P_{i2}^T & P_{i3} \end{bmatrix} > 0 \tag{31}$$

  $P_{i1} \in \mathbb{R}^{n\times n}$, $P_{i2} \in \mathbb{R}^{n\times q}$ *and* $P_{i2} \in \mathbb{R}^{q\times q}$, *for* $i = 0,\dots,m$,

- $M \in \mathbb{R}^{(n+p)\times(n+p)}$, *such that*

$$M = \begin{bmatrix} M_{11} & S_2 M_{22} \\ S_1 M_{11} & M_{22} \end{bmatrix} \tag{32}$$

  where $M_{11} \in \mathbb{R}^{n\times n}$ *and* $M_{22} \in \mathbb{R}^{q\times q}$. $S_1 \in \mathbb{R}^{q\times n}$ *and* $S_2 \in \mathbb{R}^{n\times q}$ *are some tuning matrices.*

- $Y_1 \in \mathbb{R}^{n\times p}$ *and* $Y_2 \in \mathbb{R}^{q\times p}$

and scalars $\rho > 0$, $c_1 > 0$, $c_2 > 0$ *and* $\gamma > 0$, *such that the LMIs* (33) *hold (see next page), for* $j = 1,\dots,2^{2m}$, *where the blocks* $\alpha^j_{(1,3)}$, $\alpha^j_{(1,4)}$, $\alpha^j_{(1,5)}$, $\alpha^j_{(2,5)}$, $\alpha_{(3,3)}$, $\alpha_{(3,4)}$ *and* $\alpha_{(4,4)}$ *are given by* (19d), (19e), (19k), (19l), (19h), (19j) *and* (19i), *respectively. The blocks* $\omega^j_{(1,1)}$, $\omega^j_{(1,2)}$, $\omega^j_{(2,2)}$, $\omega^j_{(2,3)}$, $\omega^j_{(2,4)}$, $\omega^j_{(1,5)}$, $\omega^j_{(1,6)}$, $\omega^j_{(2,5)}$ *and* $\omega^j_{(2,6)}$ *are as follows*

$$\begin{aligned}
\omega^j_{(1,1)} &= \sum_{i=1}^{m} \sigma_{m+i} P_{i1} + \sum_{i=0}^{m} M_{11} A_i \sigma_i + \sum_{i=0}^{m} A_i^T M_{11}^T \sigma_i \\
&\quad - Y_1 C - C^T Y_1^T - S_2 Y_2 C - C^T Y_2^T S_2^T \\
&\quad + \sum_{i=0}^{m} M_{11} A_{d_i} \sigma_i + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T \sigma_i \\
&\quad + \frac{b_1^2}{c_1} I_n + 2\frac{b_2^2 b_3^2 + b_2 b_3 b_g}{c_2} I_n \\
\omega^j_{(1,2)} &= \sum_{i=1}^{m} \sigma^j_{m+i} P_{i2} + \rho M_{11} G + \sum_{i=0}^{m} A_i^T \sigma_i^j M_{11}^T S_1^T \\
&\quad + \sum_{i=0}^{m} A_{d_i}^T M_{11}^T S_1^T \sigma_i^j - C^T Y_2^T - C^T Y_1^T S_1^T \\
\omega^j_{(2,2)} &= \sum_{i=1}^{m} \sigma_{m+i} P_{i3} + \rho S_1 M_{11} G + G^T M_{11}^T S_1^T \\
&\quad + 2\frac{b_g^2 + b_2 b_3 b_g + \rho^2}{c_2} I_q \\
\omega^j_{(2,3)} &= \sum_{i=0}^{m} \sigma_i^j P_{i2}^T - S_1 M_{11} + \rho G^T M_{11}^T \\
\omega^j_{(2,4)} &= \sum_{i=0}^{m} \sigma_i^j P_{i3} - M_{22} + \rho G^T M_{11}^T S_1^T
\end{aligned}$$

$$\begin{bmatrix} \omega^j_{(1,5)} & \omega^j_{(1,6)} \\ \omega^j_{(2,5)} & \omega^j_{(2,6)} \end{bmatrix} = P(\sigma^j)$$

The observer gains are expressed by

$$L_x = M_{11}^{-1} Y_1 \tag{34a}$$

$$L_\theta = M_{22}^{-1} Y_2 \tag{34b}$$

Proof. We rewrite the error system described by (30) in a descriptor form as follows

$$\begin{aligned}
\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{e}(t) \\ \dot{\varepsilon}(t) \end{bmatrix} &= \begin{bmatrix} 0 & I \\ H(u) + \sum_{i=0}^{m} H_{d_i} & -I \end{bmatrix} \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix} \\
&\quad + \sum_{i=0}^{m} \begin{bmatrix} 0 \\ H_{d_i} \end{bmatrix} \int_t^{t-\tau_i} \varepsilon(k)dk + \begin{bmatrix} 0 \\ H_\ell \end{bmatrix} + \begin{bmatrix} 0 \\ \overline{G}\overline{H}_g \end{bmatrix}
\end{aligned} \tag{35}$$

in order to put it in the Lyapunov function candidate $V(e)$, which will be chosen as in the first part, under the form (21). A computation of the derivative of the Lyapunov function $V(e)$ yields to

$$\begin{aligned}
\dot{V}(e) &= \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix}^T \begin{bmatrix} \overline{W}_{(1,1)} & W_{(1,2)} \\ * & -M - M^T \end{bmatrix} \begin{bmatrix} e(t) \\ \varepsilon(t) \end{bmatrix} + \sum_{i=0}^{m} \beta_i(t) \\
&\quad + \alpha_1(e) + \alpha_2(e) - \frac{1}{\gamma} \sum_{i=0}^{m} \int_t^{t-\tau_i} \dot{e}^T(s)\dot{e}(s)ds
\end{aligned}$$

where $\beta_i(t)$ is defined by (24) and

$$\begin{aligned}
\overline{W}_{(1,1)} &= \dot{P}(u) + MH(u) + \sum_{i=0}^{m} MH_{d_i}(u) + H^T(u)M^T \\
&\quad + \sum_{i=0}^{m} H_{d_i}^T(u) M^T
\end{aligned} \tag{36a}$$

$$W_{(1,2)} = P(u) - M + H^T(u)M^T + \sum_{i=0}^{m} H_{d_i}^T M^T \tag{36b}$$

$$\alpha_1(e) = \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u) \begin{bmatrix} 0 \\ H_\ell \end{bmatrix} + \begin{bmatrix} 0 \\ H_\ell \end{bmatrix}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \tag{36c}$$

$$\alpha_2(e) = \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u) \begin{bmatrix} 0 \\ \overline{G}\overline{H}_g \end{bmatrix} + \begin{bmatrix} 0 \\ (\overline{G}\overline{H}_g) \end{bmatrix}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \tag{36d}$$

However, to give an upper bound to the derivative of the Lyapounov function, we had to give an upper bound to some terms. For the term $\beta_i(t)$, an upper bound was given by the inequality (25). So, we proceed, in the sequel, by giving an upper bound to the terms $\alpha_1(e)$ and $\alpha_2(e)$.

Using inequality (11), $\alpha_1(e)$ can be upper bounded as follows

$$\begin{aligned}
\alpha_1(e) &\le \frac{1}{c_1} \begin{bmatrix} 0 \\ H_\ell \end{bmatrix}^T \begin{bmatrix} 0 \\ H_\ell \end{bmatrix} + c_1 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \\
&= \frac{1}{c_1} H_\ell^T H_\ell + c_1 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix}
\end{aligned}$$

However the product $H_\ell^T H_\ell$ can be majorated by the following expression using (1) and (5)

$$\begin{aligned}
H_\ell^T H_\ell &= [\ell(x,u) - \ell(\hat{x},u)]^T [\ell(x,u) - \ell(\hat{x},u)] \\
&\le b_1^2 e_x^T(t) e_x(t)
\end{aligned}$$

$$\begin{bmatrix}
\omega^j_{(1,1)} & \omega^j_{(1,2)} & \alpha^j_{(1,3)} & \alpha^j_{(1,4)} & \omega^j_{(1,5)} & \omega^j_{(1,6)} & M_{11} & S_2M_{22} & 0 & M_{11}G & 0 & \alpha^j_{(1,5)} \\
* & \omega^j_{(2,2)} & \omega^j_{(2,3)} & \omega^j_{(2,4)} & \omega^j_{(2,5)} & \omega^j_{(2,6)} & S_1M_{11} & M_{22} & 0 & S_1M_{11}G & 0 & \alpha^j_{(2,5)} \\
* & * & \alpha_{(3,3)} & \alpha_{(3,4)} & 0 & 0 & M_{11} & S_2M_{22} & 0 & M_{11}G & 0 & \alpha^j_{(1,5)} \\
* & * & * & \alpha_{(4,4)} & 0 & 0 & S_1M_{11} & M_{22} & 0 & S_1M_{11}G & 0 & \alpha^j_{(2,5)} \\
* & * & * & * & \frac{-1}{c_1}I_n & 0 & \dots & & \dots & & \dots & 0 \\
* & * & * & * & * & \frac{-1}{c_1}I_q & 0 & \dots & & & \dots & 0 \\
* & * & * & * & * & * & \frac{-1}{c_1}I_n & 0 & \dots & & \dots & 0 \\
* & * & * & * & * & * & * & \frac{-1}{c_1}I_q & 0 & \dots & \dots & 0 \\
* & * & * & * & * & * & * & * & \frac{-1}{c_2}I_{n+q} & 0 & \dots & 0 \\
* & * & * & * & * & * & * & * & * & \frac{-1}{c_2}I_n & 0 & 0 \\
* & * & * & * & * & * & * & * & * & * & \frac{-1}{c_2}I_q & 0 \\
* & * & * & * & * & * & * & * & * & * & * & \frac{-1}{\gamma}I_k
\end{bmatrix} < 0 \quad (33)$$

then,

$$\alpha_1(e) \le e^T(t)\mathbb{N}_1 e(t) + c_1 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \quad (37)$$

where

$$\mathbb{N}_1 = \begin{bmatrix} \frac{b_1^2}{c_1}I_n & 0 \\ 0 & 0_{q\times q} \end{bmatrix} \quad (38)$$

We note $\overline{\overline{G}} = \begin{bmatrix} 0 & 0 \\ 0 & \overline{G} \end{bmatrix}$, and we give an upper bound to α_2 as follows:

$$\begin{aligned}
\alpha_2(e) &\le \frac{1}{c_2}\begin{bmatrix} 0 \\ \overline{H}_g \end{bmatrix}^T \begin{bmatrix} 0 \\ \overline{H}_g \end{bmatrix} + c_2 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)\overline{\overline{G}}\,\overline{\overline{G}}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \\
&= \frac{1}{c_2}\overline{H}_g^T\overline{H}_g + c_2 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)\overline{\overline{G}}\,\overline{\overline{G}}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix}
\end{aligned}$$

where

$$\overline{H}_g^T\overline{H}_g = \|\overline{H}_g\|^2 \le 2\left(\|H_g\|^2 + \rho^2 e_\theta^T(t)e_\theta(t)\right)$$

with

$$\begin{aligned}
H_g^T H_g &= \left(g(x,u)\theta - g(\hat{x},u)\hat{\theta}\right)^T \left(g(x,u)\theta - g(\hat{x},u)\hat{\theta}\right) \\
&\le \left((g(x,u)\theta - g(\hat{x},u)\theta) + \left(g(\hat{x},u)\theta - g(\hat{x},u)\hat{\theta}\right)\right)^T \\
&\times \left((g(x,u)\theta - g(\hat{x},u)\theta) + \left(g(\hat{x},u)\theta - g(\hat{x},u)\hat{\theta}\right)\right) \\
&\le b_2^2 b_3^2 e_x^T(t)e_x(t) + b_g^2 e_\theta^T(t)e_\theta \\
&+ b_2 b_3 b_g\left(\|e_x^T(t)\|\,\|e_\theta(t)\| + \|e_\theta^T(t)\|\,\|e_x(t)\|\right) \\
&\le b_2^2 b_3^2 e_x^T(t)e_x(t) + b_g^2 e_\theta^T(t)e_\theta \\
&+ b_2 b_3 b_g(e_x^T(t)e_x(t) + e_\theta^T(t)e_\theta)
\end{aligned}$$

The latter inequalities lead to

$$\alpha_2(e) \le 2e^T(t)\mathbb{N}_2 e(t) + c_2 \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T F(u)\overline{\overline{G}}\,\overline{\overline{G}}^T F^T(u) \begin{bmatrix} e \\ \varepsilon \end{bmatrix} \quad (39)$$

where

$$\mathbb{N}_2 = \begin{bmatrix} \frac{b_2^2 b_3^2 + b_2 b_3 b_g}{c_2}I_n & 0 \\ 0 & \frac{b_g^2 + b_2 b_3 b_g + \rho^2}{c_2}I_q \end{bmatrix} \quad (40)$$

Hence,

$$\begin{aligned}
\dot{V}(e) \le \begin{bmatrix} e \\ \varepsilon \end{bmatrix}^T \Bigg(&\begin{bmatrix} \overline{W}_{(1,1)} + \mathbb{N}_1 + 2\mathbb{N}_2 & W_{(1,2)} \\ * & -M - M^T \end{bmatrix} \\
&+ c_1 F(u)F^T(u) + c_2 F(u)\overline{\overline{G}}\,\overline{\overline{G}}^T F^T(u) \\
&+ \gamma \sum_{i=0}^{m} F(u) \begin{bmatrix} 0 \\ H_{d_i}(u) \end{bmatrix} \begin{bmatrix} 0 \\ H_{d_i}(u) \end{bmatrix}^T F^T(u) \Bigg) \begin{bmatrix} e \\ \varepsilon \end{bmatrix}
\end{aligned}$$

Applying the Schur complement on the latter inequality, leads to inequality (41) (see next page), where the block $W_{(1,1)}$ is described by

$$\begin{aligned}
W_{(1,1)} = \dot{P}(u) + MH(u) + H^T(u)M^T + \sum_{i=0}^{m} MH_{d_i}(u) \\
+ \sum_{i=0}^{m} H_{d_i}^T(u)M^T + \mathbb{N}_1 + 2\mathbb{N}_2
\end{aligned}$$

$W_{(1,2)}$, $\mathbb{N}_1$ and $\mathbb{N}_2$ are given by (36b), (38) and (40), respectively.

Finally, using notations (27) and the information on the inputs and their derivatives, we compute inequality (41), to extract the observer gains via LMIs, which completes the proof. □

5 Discussion

1. In the literature, two major approaches have been developed to tackle the design of adaptive observer. These approaches are essentially based on:

   (a) the elaboration of a parameter adaptation law. Here, the unknown parameter vector is deduced from the stability analysis of a state observer and the convergence property of the parameter error is obtained by a persistence of excitation type constraint. Many contributions deal with this approach as in [5], [15], [17], etc.

$$\begin{bmatrix} W_{(1,1)} & W_{(1,2)} & P(u) & M & 0 & M\overline{G} & MH_{d_0} & MH_{d_1} & \dots & MH_{d_m} \\ * & -M-M^T & 0 & M & 0 & M\overline{G} & MH_{d_0} & MH_{d_1} & \dots & MH_{d_m} \\ * & * & -c_1^{-1}I_{n+q} & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ * & * & * & -c_1^{-1}I_{n+q} & 0 & 0 & 0 & 0 & \dots & 0 \\ * & * & * & * & -(c_2)^{-1}I_{n+q} & 0 & 0 & 0 & \dots & 0 \\ * & * & * & * & * & -(c_2)^{-1}I_{n+q} & 0 & 0 & \dots & 0 \\ * & * & * & * & * & * & -\gamma^{-1}I_{n+q} & 0 & \dots & 0 \\ * & * & * & * & * & * & * & -\gamma^{-1}I_{n+q} & \ddots & \vdots \\ * & * & * & * & * & * & * & * & \ddots & 0 \\ * & * & * & * & * & * & * & * & * & -\gamma^{-1}I_{n+q} \end{bmatrix} < 0 \tag{41}$$

(b) an augmented system for which the adaptive observer design is elaborated. In this case, the system dynamics are augmented with the dynamics of its unknown parameters as in [20], [26], [27], [21].

In our work, the design of our adaptive observers are based on the approach described in item (b). However, those results are established by assuming a Lyapunov function where the derivative depends of both the state and parameters errors. So, the assumption of persistent excitation is not required in our work since the matrix appearing in the derivative of the Lyapunov function is not block diagonal, unlike in [15] where the boundedness of this derivative depends only of the state error terms.

2. The use of a descriptor approach and augmenting the estimation error system as done in the present work give more additional degrees of freedom to the problem resolution and allow to overcome the problem of the product between the Lyapunov matrix $P(u)$ and the system's dynamic matrix $H(u)$ (see [28]).

3. First, due to the form of the block diagonal terms $\alpha_{(3,3)}$ and $\alpha_{(4,4)}$ appearing in both LMIs (18) and LMIs (33), the matrices M_{11} and M_{22} should be nonsingular matrices to satisfy the LMIs constraints. Notice that adding and subtracting the term $\rho G e_\theta(t)$ from equation (29a) allow to have the matrix M_{11} in the term $\omega_{(2,2)}$ in the LMI (33) given in theorem 2.

   Second, the matrix M is chosen with the form given by (17) for the following reasons:

   (a) If the matrix M is chosen under the following form

   $$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

   where $M_{11} \in \mathbb{R}^{n\times n}$ and $M_{22} \in \mathbb{R}^{q\times q}$, $M_{12} \in \mathbb{R}^{n\times q}$ and $M_{21} \in \mathbb{R}^{q\times n}$, we will come across the following problem

   $$\begin{bmatrix} Y_{11} \\ Y_{21} \end{bmatrix} = \begin{bmatrix} M_{11} \\ M_{21} \end{bmatrix} L_x$$

   $$\begin{bmatrix} Y_{12} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} M_{12} \\ M_{22} \end{bmatrix} L_\theta$$

   Then, the existence of the observer's gains depends on the following rank conditions

   $$\operatorname{rank}\begin{bmatrix} M_{11} \\ M_{21} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} M_{11} & Y_{11} \\ M_{21} & Y_{21} \end{bmatrix}$$
   $$\operatorname{rank}\begin{bmatrix} M_{12} \\ M_{22} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} M_{12} & Y_{12} \\ M_{22} & Y_{22} \end{bmatrix}$$

   which add some non-convex constraints to satisfy in theorem 1 and 2.

   (b) Putting the matrix M with a diagonal form, i.e. $M_{12} = 0$ and $M_{21} = 0$, implies that the blocks $\alpha_{(2,2)}$ and $\omega_{(2,2)}$ in LMIs (18) and (33) respectively, will be written as follow

   $$\alpha^j_{(2,2)} = \sum_{i=1}^{m} \sigma^j_{m+i} P_{i3}$$
   $$\omega^j_{(2,2)} = \sum_{i=1}^{m} \sigma_{m+i} P_{i3} + \frac{b_g^2 + b_2 b_3 b_g + \rho^2}{c_2} I_q$$

   and one can see that the LMIs (18) and (33) can not be satisfied.

   So, to avoid the above rank constraints, we set $M_{21} = S_1 M_{11}$ and $M_{12} = S_2 M_{22}$ where matrices S_1 and S_2 are a priori chosen tuning parameters. This leads to $Y_{21} = S_1 Y_{11}$ and $Y_{12} = S_2 Y_{22}$.

   However, one can see that, unlike matrix S_1, matrix S_2 does not appear in the diagonal blocks $\alpha_{(2,2)}$ and $\omega_{(2,2)}$ of LMIs (18) and (33), respectively. By the way, we can set $S_2 = 0$. Whereas, we need the condition $S_1 \neq 0$.

4. The obtained results in theorem 2 may appear as an extension of theorem 1. However, it is not the case. For the observer design in section 3, the nonlinear Lipschitz functions are taken as

   $$\ell(x,u) = 0 \ \text{ and } \ g(x,u) = 1$$

   which imply that assumption 2 will be as follows

$$b_1 = 0, \quad b_2 = 0 \text{ and } b_g = 1$$

Applying this assumption on the results of theorem 2, do not lead to the results obtained in theorem 1, due to the fact that this assumption does not cancel some terms in the block $\omega^{j}_{(2,2)}$ of the LMIs (33). In addition, the bound b_3 of the unknown parameters θ is not required in the design of adaptive observer in section 3.

6 Numerical examples

To illustrate the efficiency and the feasibility of our results, we give in the sequel some numerical examples. Let us consider a bilinear time delay system, with

$$A_0 = \begin{bmatrix} -4 & 2 \\ -1 & -1.52 \end{bmatrix} \quad A_1 = \begin{bmatrix} -0.4 & 0.8 \\ 0.27 & -0.4 \end{bmatrix}$$
$$A_{d_0} = \begin{bmatrix} -0.1 & 0.01 \\ -0.11 & -0.68 \end{bmatrix} \quad A_{d_1} = \begin{bmatrix} -0.8 & 0.08 \\ -0.2 & -0.04 \end{bmatrix}$$

We assume that the system is controlled by one bounded input control $u(t) = u_1(t)$

$$-0.2 < u(t) = u_1(t) = 0.2\sin(0.1t) < 0.2$$

with

$$B = \begin{bmatrix} 0.8 \\ 0.01 \end{bmatrix}$$

One can see, that the derivative is also bounded, such that

$$-0.02 < \dot{u}(t) < 0.02$$

Then, we assume that the system is affected by one constant unknown parameters θ, where

$$G = \begin{bmatrix} -0.09 \\ 0.9 \end{bmatrix}$$

The time delays are constant and known such that

$$\tau_0 = 0.9, \text{ and } \tau_1 = 0.4$$

The available measurement vector is given by

$$y(t) = x_1(t)$$

The Lipschitz nonlinear function $g(x,u)$ is bounded and chosen as follows

$$-0.2 \leq g(x,u) = 0.2\sin(u(t) - x_1(t)) \leq 0.2$$

and the function $\ell(x,u)$ is given by

$$\ell(x,u) = \begin{bmatrix} \sin(x_1(t)e^{-0.2u(t)}) \\ \sin(x_2(t)e^{-0.2u(t)}) \end{bmatrix}$$

The unknown parameters θ will be assumed to be bounded by $b_3 = 0.5$ only for the simulation in section 6.2.

6.1 Numerical results related to the first observer design without Lipschitz nonlinearities

By computing the LMIs (18), given in theorem 1, in the set of the vertices of the convex polytope $\mathcal{P}$, using the toolbox "Lmilab", we obtain the following matrices

$$P_0 = \begin{bmatrix} 8.1168 & -1.8823 & -1.7727 \\ -1.8823 & 7.9346 & -3.7539 \\ -1.7727 & -3.7539 & 8.8302 \end{bmatrix}$$
$$P_1 = \begin{bmatrix} 1.1724 & -1.0393 & -0.37411 \\ -1.0393 & 1.1054 & 0.36291 \\ -0.37411 & 0.36291 & 0.081736 \end{bmatrix}$$
$$M_{11} = \begin{bmatrix} 2.3048 & 0.15176 \\ 0.15176 & 2.4665 \end{bmatrix}$$
$$M_{22} = 7.1197$$
$$S_1 = \begin{bmatrix} -0.9 & -1.12 \end{bmatrix}$$
$$S_2 = \begin{bmatrix} -0.003 \\ 0.02 \end{bmatrix}$$
$$Y_1 = \begin{bmatrix} -2.8248 \\ -3.067 \end{bmatrix}$$
$$Y_2 = 5.6283$$
$$\gamma = 0.1$$

which yield to the following observer's gains

$$L_x = M_{11}^{-1} Y_1 = \begin{bmatrix} -1.1484 \\ -1.1728 \end{bmatrix}$$
$$L_\theta = M_{22}^{-1} Y_2 = 0.79053$$

Figures 1, 2 and 3 show that the observer gives a suitable estimation of the states x_1 and x_2 and the unknown parameters θ. They prove also that even if the values of the unknown parameters change, but still constant, the estimation of the states and the unknown parameter converge to their real values.

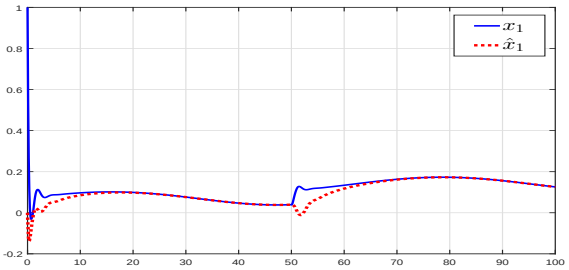


Figure 1: Variation of the state x_1 (—) and its estimation $\hat{x}_1$ ($\cdots$).

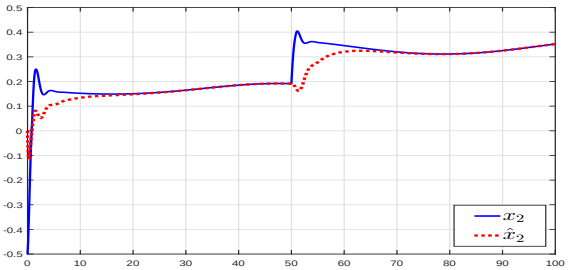


Figure 2: Variation of the state x_2 (—) and its estimation $\hat{x}_2$ ($\cdots$).

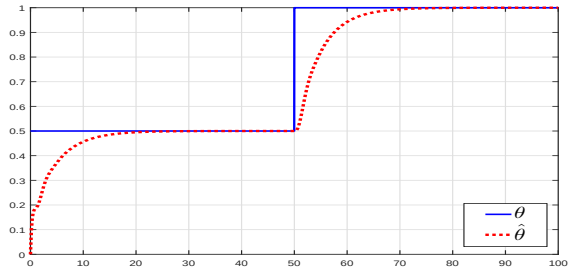


Figure 3: Variation of the unknown parameter θ (—) and its estimation $\hat{\theta}$ ($\cdots$).

6.2 Numerical results related to the second observer design with Lipschitz nonlinearities

By setting the tuning matrices S_1 and S_2 as follows

$$S_1 = \begin{bmatrix} -0.21963 & -0.36278 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 0.11754 \\ 0.049037 \end{bmatrix}$$

and fixing $b_1 = 0.2$, $b_2 = 0.2$ and $\rho = 0.1$, the resolution of the LMIs (33) given in theorem 2 using the toolbox "sdpt3" of Matlab, leads to the following observers gains

$$L_x = M_{11}^{-1} Y_1 = \begin{bmatrix} 0.41697 \\ -1.0876 \end{bmatrix}$$

$$L_\theta = M_{22}^{-1} Y_2 = 1.0292$$

with

$$P_0 = \begin{bmatrix} 1.1844 & -0.28959 & -0.032828 \\ -0.28959 & 0.61048 & 0.015679 \\ -0.032828 & 0.015679 & 0.084515 \end{bmatrix}$$

$$P_1 = \begin{bmatrix} 0.22818 & -0.07406 & -0.024134 \\ -0.07406 & 0.10067 & -0.0056328 \\ -0.024134 & -0.0056328 & -0.00030159 \end{bmatrix}$$

$$M_{11} = \begin{bmatrix} 0.31247 & 0.030714 \\ 0.030714 & 0.4203 \end{bmatrix}$$

$$M_{22} = 0.19653$$

$$Y_1 = \begin{bmatrix} 0.096884 \\ -0.44431 \end{bmatrix}$$

$$Y_2 = 0.20227$$

$$c_1 = 1.12, \quad c_2 = 0.58, \quad \gamma = 0.054$$

Using the obtained gains into the observer yields to a suitable estimation of the states $x(t)$ and the unknown parameter θ. The following figures show the effectiveness of our design.

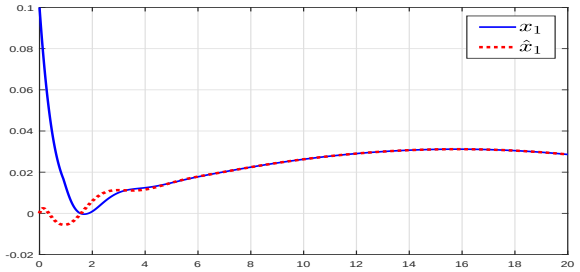


Figure 4: Variation of the state x_1 (—) and its estimation $\hat{x}_1$ ($\cdots$).

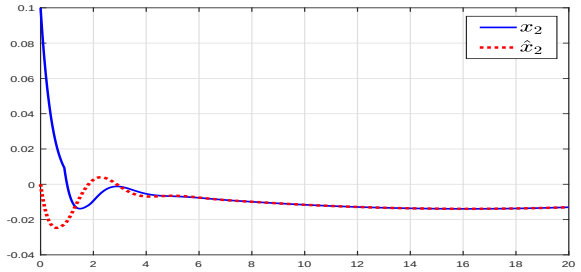


Figure 5: Variation of the state x_2 (—) and its estimation $\hat{x}_2$ ($\cdots$).

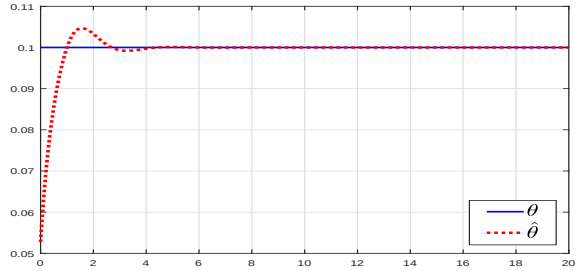


Figure 6: Variation of the unknown parameter θ (—) and its estimation $\hat{\theta}$ ($\cdots$).

7 Conclusion

The proposed adaptive observer for a class of nonlinear time delay system developed in this paper allows a suitable estimation of the state and the unknown parameter vectors, simultaneously. In a first time, a bilinear system with time delays is considered, which was extended by the addition of some nonlinear Lipshitz functions. The observer gains are obtained by solving LMIs in the vertices of a convex polytope. The simulation results show the performances and feasibility of the proposed approach. An extension of our adaptive observer design was presented in a second time, when a nonlinear Lipschitz functions interfere in the dynamics of a bilinear time delay system.

References

[1] A. Sassi, H. Souley Ali, M. Zasadzinski, and K. Abderrahim, "Full order adaptive observer design for time delay bilinear system." *In the Proceedings of the 6th International Conference on Systems and Control*, pp. 267–272, June 2017.

[2] G. Bastin and M. Gevers, "Stable adaptive observers for nonlinear time-varying systems." *IEEE Transactions on Automatic Control*, vol. 33, no. 7, pp. 650–658, 1988.

[3] G. Besançon, "Remarks on nonlinear adaptive observer design." *Systems and Control Letters*, vol. 41, pp. 271–280, 2000.

[4] Y. Choi, H. Shim, Y. Son, and J. Seo, "Design of an adaptive observer for a class of nonlinear systems," *International Journal of Control, Automation and Systems*, vol. 1, no. 1, pp. 28–34, 2003.

[5] M. Ekramian, F. Sheikholeslam, S. Hosseinnia, and M. Yazdanpanah, "Adaptive state observer for Lipschitz nonlinear systems," *Systems and Control Letters*, vol. 62, pp. 319–323, 2013.

[6] L. Xie and P. Khargonekar, "Lyapunov-based adaptive state estimation for a class of nonlinear stochastic systems," *Automatica*, vol. 48, pp. 1423–1431, 2012.

[7] L. Zhao, X. Li, , and P. Li, "Adaptive observer design for a class of mimo nonlinear systems," *Proceedings of the 10th Wold Congress on Intelligent Control and Automation*, pp. 2198–2203, 2012.

[8] S. Ibrir, W. Xie, and M. C. Su, "Observer-based control of discrete-time Lipschitizian nonlinear systems: Application to one-link flexible joint robot." *International Journal of Control*, vol. 78, no. 6, pp. 385–395, 2005.

[9] A. Zemouche and M. Boutayeb, "On LMI conditions to design observers for Lipschitz nonlinear systems." *Automatica*, vol. 49, pp. 585–591, 2013.

[10] J. Richard, "Time-delay systems : An overview of some recent advances and open problems," *Automatica*, vol. 39, no. 10, pp. 1667–1694, 2003.

[11] M. Kadhraoui, M. Ezzine, H. Messaoud, and M. Darouach, "Design of unknown input functional observers for delayed singular systems with state variable time delay," *Transactions on Systems and Control*, vol. 10, pp. 503–509, 2015.

[12] C. Briat, O. Sename, and J. Lafay, "Design of LPV observers for LPV time-delay systems: an algebraic approach." *International Journal of Control*, vol. 84, no. 9, pp. 1533–1542, 2011.

[13] E. Fridman, "Tutorial on Lyapunov-based methods for time-delay systems," *European Journal of Control*, vol. 20, pp. 271–283, 2014.

[14] K. Gu, V. Kharitonov, and J. Chen, *Stability of Time-Delay Systems*. Birkhäuser Boston, 2003.

[15] Y. Cho and R. Rajamani, "A systematic approach to adaptive observer synthesis for nonlinear systems," *IEEE Transactions on Automatic Control*, vol. 42, no. 4, pp. 534–537, 1997.

[16] H. Wang, Y. Dong, and W. Qin, "Adaptive observer design for a class of Lipschitz nonlinear systems," *In the proceedings of the 30th Chinese Control Conference*, pp. 665–669, July 2011.

[17] M. Pourgholi and V. Majd, "Robust adaptive observers design for lipschitz class of nonlinear systems." *International Journal of Electrical and Computer Engineering*, vol. 6, no. 3, pp. 275–279, 2012.

[18] Q. Zhang and B. Deylon, "A new approach to adaptive observer design for mimo systems," *American Control Conference*, vol. 2, pp. 1545 – 1550, 2001.

[19] G. Besançon, J. D. León-Morales, and O. Huerta-Guevara, "On adaptive observers for state affine systems." *International Journal of Control*, vol. 79, no. 6, pp. 581 – 591, 2010.

[20] M. Alma and M. Darouach, "Adaptive observer design for a class of linear descriptor systems." *Automatica*, vol. 50, pp. 578–583, 2014.

[21] M. Farza, I. Bouraoui, T. Ménard, and R. Ben Abdennour, "Adaptive observer for a class of uniformly observable systems with nonlinear parametrization and sampled outputs." *Automatica*, vol. 50, pp. 2951–2960, 2014.

[22] M. Alma, H. S. Ali, M. Darouach, and N. Gao, "An H_∞ adaptive observer design for linear descriptor systems." *2015 American Control Conference*, pp. 4838–4843, May 2015.

[23] A. Sassi, H. Souley Ali, S. Bedoui, K. Abderrahim, and M. Zasadzinski, "Design of a functional adaptive observer for bilinear delayed systems." *13th IFAC workshop on Time Delay Systems TDS*, vol. 49, no. 10, pp. 25–30, 2016.

[24] A. Sassi, M. Zasadzinski, H. Souley Ali, and K. Abderrahim, "Design of an H-infinity adaptive observer for bilinear delayed systems." *The 20th World Congress of the International Federation of Automatic Control (IFAC WC)*, vol. 50, no. 1, pp. 2923–2928, July 2017.

[25] H. Khalil, *Nonlinear Systems, Third Edition*. USA: Prentice Hall, 2002.

[26] Q. Zhang, "Revisiting different adaptive observers through a unified formulation." *Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference.*, pp. 3067–3072, December 2005.

[27] T. Ahmed Ali, R. Postoyan, and F. Lamnahbi Lagarrigue, "Continuous-discrete adaptive observers for state affine systems." *Automatica*, vol. 45, pp. 2986–2990, 2009.

[28] B. Gérard, H. Souley Ali, M. Zasadzinski, and M. Darouach, "H-infinity filter for bilinear systems using LPV approach," *IEEE Transactions on Automatic Control*, vol. 55, no. 7, pp. 1668–1674, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 384-393 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

A Joint Safety and Security Analysis of message protection for CAN bus protocol

Luca Dariz[*,1], Gianpiero Costantino[2], Massimiliano Ruggeri[1], Fabio Martinelli[2]

[1]*CNR-IMAMOTER, Via Canal Bianco 28, Ferrara, Italy*

[2]*CNR-IIT, Via G. Moruzzi, 1, Pisa, Italy*

ARTICLEINFO

Article history:
Received: 15 November, 2017
Accepted: 10 January, 2018
Online: 10 February, 2018

Keywords:
Automotive
CAN bus
Security by design
Security-properties
Safety

ABSTRACT

One of the prominent challenges of the automotive-transportation system is represented by the integration of security and safety properties within protocols, applications and connectivity mechanisms. A joint safety/security design can sometimes expose to trade-offs, since their requirements may not match perfectly or even be incompatible. This paper analyses an example of security and safety design, by combining integrity with encryption considering the constraints of a typical CAN protocol and real-time traffic. The analysis is presented modelling differently attackers, packet fragmentation issues and the residual probability of error of the combined scheme.

1 Introduction

The common thought about cars is that they are mechanical devices employed by passenger to move from a place to another. This is not true any more, or at least is not completely true since cars, or in general vehicles, in the last lustrum evolved to offer several services and connections that turn them into Cyber-Physical Systems (CPS) by the combination of sensing/actuation, processing, storing, and networking capabilities, as described by Fortino et al. [1]. Applications, sensors, park and driving assistants are integrated into commercial vehicles and they are considered standard features present in entry-level model of cars. Features, such as Internet connectivity, enlarge the attack surface of vehicles and, in particular, traditional communication protocols developed to work on isolated environments could not maintain the same level of robustness when new variables are taken into account. This is especially true if one considers that connectivity itself can be provided in multiple heterogeneous ways, which are not always predictable by the manufacturer anymore. For example, in [2], the author proposed a connectivity solution based on smartphones. On the other hand, practically all Electronic Control Units (ECUs) on a car are connected to one or more CAN busses, which is the traditional interface for intra-vehicle communications. The CAN bus use messages whose payload is at maximum 64bits length, and depending on the payload set, it can enable specific functionality of the vehicle, for instance enabling the accelerator of a car. The CAN bus protocol was not designed to embed security properties such as: *Authentication*, *Integrity* and *Confidentiality*, and the security aspects are left to higher layers of the protocols-stack, for instance the application level. An example of attack on the CAN bus protocol is that performed on a Jeep Cherokee by Valasek and Miller in 2015 [3], where the authors showed how to hack and remotely control a Jeep Cherokee. This attack exploit a security flaw in the *In-vehicle Infotainment* (IVI) system of the car to access the CAN bus network of the car. To fix this flaw, the Fiat Chrysler was forced to push a software update [4].

Security issues are a major challenge for connected vehicles, as reported in [5], and this is recognised also to have a relevant impact on vehicle's safety [6]; it is clear that *security* and *safety* are fundamental in the design of an intelligent transportation system, especially regarding the communication protocols and message protection schemes. It is important to note that the combination of safety and security issues is not exclusive of vehicular systems, but it affects the whole IoT world, especially systems which functionality is considered critical (e.g. medical devices).

In this work, we propose a solution to turn the CAN bus protocol as a *Security by Design* protocol by integrating authentication, integrity and confidential-

[*]Corresponding Author: Luca Dariz, CNR-IMAMOTER, Via Canal Bianco 28, Ferrara, Italy

https://dx.doi.org/10.25046/aj030147

ity properties. Our solution proposes the adoption of a Message Authentication Code (MAC), targeted for CAN bus messages, that is then encrypted with an additional key. The message created guarantees authentication and integrity through the MAC, and confidentiality with the additional encryption. Our defence strategy is studied to be applied against a model of attacker that runs both a Honest-But-Curious (HBC) or Fully Malicious attack strategy. Furthermore, our solution is evaluated from a safety point of view, in particular regarding the residual probability of error (P_{re}). This is necessary since the outcome of the security MAC i.e. accept or reject a particular message is a form of error detection which could reveal also transmission errors (e.g. caused by noise), and the message containing the MAC could bring safety-critical information. For example, this scheme could be applied to SAE J1939 Torque/Speed Control 1 message, which embeds a 4-bit checksum within the 8-byte CAN payload. The residual probability of error is first evaluated using an ideal block cipher model, then simulation results are presented for a specific implementation choice. We show that averaging over the secret keys and over the possible messages, the value of P_{re} depends only on the length of the integrity tag used to decide whether the message is valid or not, and this is *independent* from how the integrity tag is generated. On the other hand, we show how the worst case combination of key and message reduces massively the ability of this scheme to detect transmission errors, assuming a simplified channel model. The worst-case P_{re} is then simulated, with considerations on the difficulty of finding the worst-case combination of key and message.

The main contributions of this paper is to analyse a message protection protocol from both a security and safety point of view, and highlight the trade-offs that result from the analysis. This paper is an extension of work originally presented in *IEEE MT-ITS* [7] and our contributions are summarized as follows:

- We improved the description of the analytical model for residual error probability by providing more details on the error model, new explanation on the computation of P_{re} and analytical results;
- We expanded the security analysis and we better modelled the attacker;
- We improved the safety analysis giving more details on the fragmentation and retransmission issues;
- We added more results and better described the analytical model depending on the statistical distribution of errors in the simple case of binary symmetric channel;
- We redraw the plots with the simulation results, using colours, to be easier to understand.

This paper is structured as follows: section 2 reviews the state of the art with regard to MAC algorithms and best practices, as well as error detection mechanisms; section 3 presents the message protection scheme discussed in this paper, discussing some design choices. In section 4, we introduce the attacker model and we make a security consideration on the MAC size. Section 5 exposes two aspects usually relevant for the safety of the system, that are packet fragmentation and the probability of residual error P_{re}. Section 6 discusses other message protection schemes and their security and safety properties, compared to the scheme proposed here. Finally, section 7 concludes the paper with some motivation for future research directions.

2 Related Works

The following works refers to lightweight Message Authentication Code solutions for devices that have limited computational resources, like processors and memory. Chowdhury and Dasbit in [8] introduce *LMAC*, a Lightweight Message Authentication Code (MAC) of 64bits dimension for Wireless Sensor Network that uses hash based symmetric key MAC. The authors show that LMAC is secure against passive and active attacks and it has a low overhead compared with other similar solutions. Another 64bits MAC, called *Chaskey* is presented by the authors of [9]. Chaskey uses 128bits key to generate a MAC which length is of 64bits or more. The authors say that Chaskey generates MAC that are suitable for 32bit Microcontroller and that it does not suffer of MAC truncation [10]. In [11] the authors show two versions of lightweight MAC of 64 and 128bits called *TuLP-64* and *TuLP-128* that are resources efficient and are though for body sensor networks. Then, in [12] the authors presented a lightweight MAC suitable for Smart Grid communications in which two devices reach mutual authentication by sharing a session key exchanged using Diffie-Hellman and a hash-based authentication code technique.

Regarding safety properties, such as the residual probability of error, Schiller and Mattes have analysed different ways of using nested CRC codes, for example in [13]. However, when it comes to cryptographic algorithms the kind of errors treated are related to security properties and possible vulnerabilities, see for example the survey in [14]. To the best of our knowledge, an explicit model for the residual probability of error of a system using cryptographic algorithms has not been developed, although the statistical properties of symmetric ciphers are sometimes studied in-depth, but always in the context of security, see for example [15].

3 Message protection scheme

The main scheme analysed in this paper is based on encryption, and is represented in Figure 1. In general, a message μ with length μ_{size} bits is combined with an integrity tag $\tau = H(k_2, \mu)$ of length τ_{size} , where k_2

is the *authentication key*. The combined $\mu\|\tau$ is then encrypted to obtain the ciphertext $c = ENC(k_1, \mu\|\tau)$, where k_1 is the *encryption key*; the ciphertext is then transmitted on the CAN bus. The receiver receives c', decrypts it and checks if $\tau' = H(k_2, \mu')$, with $\mu'\|\tau' = DEC(k_1, c')$, to decide if the message is valid.

There can be some variations in this scheme; for example the integrity code can be appended to μ to form the plaintext (also known as MAC-then-encrypt approach) with $c = ENC(k_1, \mu\|\tau)$, or it can be excluded from encryption (encrypt-then-MAC approach) with $c = ENC(k_1, \mu)\|\tau$. In both cases the MAC has to be computed using the original message μ. Sometimes the MAC-then-encrypt approach is considered less secure, for example see [16], but in this paper the scheme has no padding and a fixed length of the message, so these considerations do not apply. Considering the CAN bus, the first approach is more practical since there exist encryption algorithms with 64-bit block size, equal to the maximum payload of a CAN message; in this case there is no need for additional data to perform the encryption, and the ciphertext is computed as $c = ENC(k_1, \mu\|\tau)$. On the other hand, if the plaintext is different from the block size (like in the encrypt-then-MAC approach), the cipher must be used in counter mode, or a stream cipher can be used. Either way, there needs to be additional information shared between the sender and the receiver, e.g. a nonce, to perform the encryption; in this case the ciphertext is computed as $c = ENC(k_1, I, \mu)\|\tau$ with I being the shared information.

Another variant is to avoid the use of two different shared keys and define the integrity code as $\tau = H(\mu)$, where $H(\cdot)$ is a hash function like SHA1 or an error-detection code of the CRC family.

In this paper we consider only different possibilities for the integrity tag $\tau = H(\mu)$, which can be a proper Message Authentication Code, a hash function or a CRC; we do not consider then the encrypt-then-MAC approach, so we can define the plaintext $m = \mu\|\tau$.

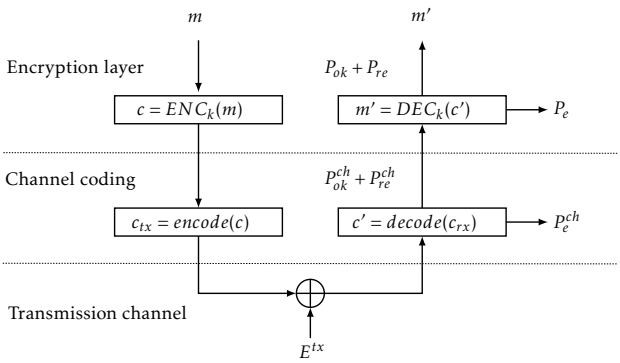


Figure 1: Message protection scheme.

[1] Explanation cited from https://www.krackattacks.com

4 Security considerations

When cars were not connected to the Internet, the attack surface was limited to local entry points that could make cars vulnerable to security attacks. Since connectivity has been embedded as feature inside cars, the attack surface is increased and it is not limited to local attacks, but also remote ones make a vehicle more vulnerable. This provides a higher impact to the security and safety of cars' passengers, and in addition, gives to attackers more and more kinds of attacks that they are able to exploit.

When, we talk about to local attacks, we refer to physical inputs to the vehicle that can be represented, for instance, by the auxiliary jack of the CD-Rom, USB ports, and the ODB2 connector. All these physical connectors require that an attacker has the opportunity to access the car and manipulate or alter the input to perform its attack. This, however, is possible but the impact is limited and less transparent to passenger. On the other side, remote attacks made on the wireless inputs enlarge the attack surface and may make attacks more relevant. Wireless inputs can be: the Bluetooth, Wi-Fi and cellular connections. In particular, exploiting the cellular connection, it would be possible to remotely access the In-vehicle infotainment (IVI) system of a car, and from that reading and writing into the CAN bus [3]. In fact, the IVI system are consoles that often run an operating system such as Windows CE, Linux, and may even run Android. So, attackers may exploit known vulnerabilities of unpatched version of those operating systems to perform their attacks. On this scenario, the Key Reinstallation Attacks (KRACKs) [17] has demonstrated that a vulnerability in the WPA2 protocol could allow *"attackers to use this novel attack technique to read information that was previously assumed to be safely encrypted. This can be abused to steal sensitive information such as credit card numbers, passwords, chat messages, emails, photos, and so on*[1]*."*. Thus, using KRACKs and exploiting an unpatched version of the Wi-FI provided by the IVI, an attacker could jump in the car's network getting sensitive information, and using that access as gateway to force and get in the CAN bus network of the vehicle.

4.1 Attacker Model

In this paper, we analyse the CAN protocol from the security and safety point of views. In particular, this section aims at modelling attackers, defining which are the attacks that they may exploit.

In the modelling phase, we consider that attackers may have local or remote access to the vehicle to compromise the CAN bus by forging or altering messages that may be considered valid by recipients. For instance, attackers may be able to exploit a weakness of the authentication module to remotely access the CAN bus network using a classic IP connection, and

once inside the vehicle, they forge valid message or even altering their contents.

To minimize the power of attackers, our defence strategy foresees that more attackers are not able to forge valid messages, keeping enabled confidentiality of proper messages generated. Also, we aim at identifying messages that were altered, i.e., *losing of integrity*. Thus, our defence strategy is based on three security properties that are:

Authentication: A recipient should be able to verify whether a message is sent by a legitimate sender;

Integrity: A recipient should be able to verify whether a message has been altered during its transmission;

Confidentiality: it guarantees that the content of a message is not revealed to an illegitimate entity, as it can happen with the Man-in-the-Middle (MITM) attack;

We apply our defence strategy against attackers who can play the following roles:

Honest-but-Curious (HBC): Also known as *Passive Attack*; an attacker may exploit the information legitimately gleaned by capturing messages exchanged over the CAN bus network, but he/she will not perform any malicious activity to harvest it.

Fully Malicious (FM): Also known as *Active Attack*; an attacker is able to forge or alter messages that are considered valid, after a verification step, by the recipient. So, the attacker strategy is to succeed in at least one of the following attacks:

- *Impersonation* attack: the attacker is able to assume the identity of one of the legitimate parties;
- *Guessing* attack: the attacker is able to forge a valid MAC after a number of trials.
- *Replay* attack: the attacker is able to re-use valid messages with a malicious or fraudulent aim;
- *Sniffing* attack: the attacker is able to read the content of any messages exchanged through the CAN bus network:

In Table 1, we combine the defence strategy to each attack presented above aiming at blocking or mitigating the corresponding attack.

4.2 Security considerations on encrypted MAC and MAC size

We introduced the Message Authentication Code in the CAN-message payloads to provide *by-design authentication* and *integrity* properties. At the same time, by encrypting $\mu||\tau$, we added the *confidentiality* property in messages to avoid that an attacker sniffs content messages exchanged among ECUs. However, even not considering encryption on messages, but even only the first two properties, an attacker is capable to sniff messages, but she will not easily forge valid messages due to the MAC. With encryption, the attacker's knowledge is still minor, and the probability to forge a valid message is linked to the *guessing attack* plus the encryption of $\mu||\tau$.

Dworkin in [18] points out the importance of choosing a robust MAC to be resistant against the guessing attack. In particular, Dworkin explains that a sound MAC is provided with a size greater than 64bits, i.e., $\tau_{size} \geq$ 64-bits. However, due to the standard CAN bus payload restriction, i.e., 64bit in total, it is very hard to keep that inequality true and we need a workaround to guarantee the security properties with a specific level of risk. Thus, the workaround can be implemented through two different strategies: i) concatenating a MAC which size is at least 64-bit, or ii) limiting the number of repeated trials of an attacker before considering invalid the key that generates the MAC. The first solution may cause the fragmentation issue that we detail better in §5.1. Instead, the second solution can be the best candidate for our workaround on τ_{size}. To this purpose, in [18], Dworkin illustrates how to calculate the right τ_{size} depending on the following two bounds:

MaxInvalids: as the limit on the number of trials that an attacker can perform before the key is retired;

Risk: the highest acceptable probability for an inauthentic message to be accepted as valid;

Then, due the above parameters, the τ_{size} should satisfy the following inequality:

$$\tau_{size} \geq lg(\frac{MaxInvalids}{Risk}). \quad (1)$$

Our goal is to satisfy the inequality 1 with a value of τ_{size} that is greater than or equal to 16. From inequality 1, the system can tolerate up to 30 (2^5) messages before considering the key invalid, and the system can accept 2^{-11}, i.e., $Risk = 2048$, chance of inauthentic messages. So, considering that the payload-size of a CAN-message is generically 48bits (μ_{size}) and the maximum bandwidth of the communication channel is 64bit ($bandwidth_{max}$), we obtain that $\mu_{size} + \tau_{size} \leq bandwidth_{max}$. The inequality 1 defines that the lowest condition to have a $\mu||\tau$ message for the CAN bus protocol is to have $MaxInvalids = 2^5$ and $Risk = 2048$.

5 Safety considerations

5.1 Fragmentation

It is well known that a standard CAN message is too short for the proper implementation of many security properties, since the maximum allowed payload length is 8 bytes. However there are use cases where using a second CAN message (e.g. for authentication

Attack	*Description*	**Defence**
Impersonation	Generating messages being identified as legitimate party	Authentication
Guessing	Forging or altering messages that are valid by the recipient	Confidentiality, Authentication, Integrity
Replay	Re-use messages that are considered valid by the recipient	Authentication
Sniffing	Read content of messages	Confidentiality

Table 1: Summary of attack and defence strategies.

purposes) is not acceptable since it introduce unnecessary latency in the complete reception of a message; this is mainly due to the low speed of CAN bus, which is typically 250 kB/s on vehicular networks, and can reach the maximum of 1 MB/s. Another reason is the increase of residual error rate in the communication (see for example the appendix D.5.2 of ISO 15998 [19]); however, this increase could be tolerated more easily since, to a first approximation, the P_{re} of a two-message scheme is roughly the double of the P_{re} of a single message. With these limitations, common application requirements allow for a rather limited security level achievable without changing the network communication protocol stack, and is fundamentally due, for a point-to-point communications, to the limited payload length of a single CAN 2.0 bus message. The use of CAN-FD could relax these limitations, but even in this case the maximum payload length is 64 bytes.

5.2 Retransmissions

In order to tolerate packet loss, communication protocols usually employ some form of re-transmission handling. Basically this means that if an error is detected on the packet, or if a timeout for packet reception expires, the transmitter sends again the same packet. Different variants of ARQ schemes (Automatic Repeat reQuest) exist, but the essential principle is that a data packet is sent more than one time without modifications. While this feature is usually handled transparently at link or transport layer, it is a potential security vulnerability, since an attacker could either inject errors or transmit duplicated packets to perform a replay attack (§4.1). Similar attacks are not hypothetical but have been demonstrated in reality, one of the most recent examples being the key reinstallation attack against WPA2 [17]. It is then desirable to handle retransmissions at the application level, if any, so sensitive security information like nonces, which must be used exactly one time, are handled properly. It is also worth considering that fragmentation makes retransmissions more complicated to handle efficiently, since here a small portion of a packet can be lost.

5.3 Probability of Residual Error

The analysis of the probability of residual error P_{re} in a communication protocol is usually a difficult task. In order to simplify the model, it is assumed that the channel coding is independent from the encryption scheme and it is possible to obtain a measure of P_{re} for channel coding, which correspond to P_{re}^{ch} with reference to Figure 1. In this way, it is possible to focus only on the performance of the message protection scheme discussed in this paper. Here P_{re} is obtained assuming Shannon's ideal cipher model, which has been used in other works like [20], and the results are compared with simulations where real algorithms are used for encryption and integrity.

5.3.1 Error model

According to Shannon's model, an ideal cipher is a random family of permutations, chosen independently for each possible key. More precisely, suppose $\mathcal{K}$ is the set of all keys and $\mathcal{M}$ is the set of all messages. An ideal block cipher is a map $ENC : \mathcal{K} \times \mathcal{M} \to \mathcal{M}$ where, for each key $k \in \mathcal{K}$, the encryption function $ENC_k(\cdot) = ENC(k, \cdot)$ is a random permutation on the message set $\mathcal{M}$ (independent of any other permutation). The same considerations apply to the decryption function $DEC_k(\cdot) = DEC(k, \cdot)$, which being the inverse of a random permutation is itself a random permutation.

In this context, it is more useful to fix a specific pair (k, m) of key and message, and consider the cipher text message set as:

$$\mathcal{C}' = \{c' : c' = c + e', e' \in \mathcal{E}'\} \tag{2}$$

while the plain text message set is:

$$\mathcal{M}' = \{m' : m' = m + e, e \in \mathcal{E}\} \tag{3}$$

where the set $\mathcal{E}'$ is the set of all possible undetected error vectors after channel decoding and before decryption, and $\mathcal{E}$ is the set of all error vectors after decryption. The addition here correspond to a bitwise XOR. While the elements of $\mathcal{E}'$ correspond basically to undetected transmission errors, the elements of $\mathcal{E}$ can be seen as undetected transmission errors transformed by the decryption function. For this reason, the distribution of the values in $\mathcal{E}$ depends on the distribution of $\mathcal{E}'$, in a different way for each different pair (k, m), since the function $E(k, \cdot)$ will correspond to a different and independent permutation. The relation between $\mathcal{E}'$ and $\mathcal{E}$ is then of the form $ENC : \mathcal{K} \times \mathcal{M} \times \mathcal{E}' \to \mathcal{E}$ which can be rewritten as the map $ENC_{(k,m)} : \mathcal{E}' \to \mathcal{E}$ identified by a specific pair of values (k, m). The explicit form can be derived from Figure 1 as

$$\begin{aligned} e = ENC_{(k,m)}(e') &= DEC(k, c') && (4)\\ &= DEC(k, c + e') && (5)\\ &= DEC(k, ENC(k, m) + e') + m && (6) \end{aligned}$$

Being $ENC(k,\cdot)$ and $DEC(k,\cdot)$ random permutations, equation 6 represent a random map from $\mathcal{E}'$ to $\mathcal{E}$. Each pair (k,m) correspond then to a different map $ENC_{(k,m)}$, independent from other maps. This means that each error vector e' on the transmission channel correspond to a random error vector e after decryption. Assuming that the pair (k,m) is chosen uniformly from the set $\mathcal{K}\times\mathcal{M}$, the random error vector e is itself uniformly distributed. In practice, the relation between e and e' is highly non-linear, since it strongly depends on the ENC and DEC functions, which in real block ciphers are usually highly non-linear functions themselves. This means that the approximation of uniform distribution of e, for any given e' and varying (k,m), is accurate as long as these conditions apply:

1. the real block cipher approximates a random permutation;
2. the messages in $\mathcal{M}$ are uniformly distributed;
3. the keys in $\mathcal{K}$ are uniformly distributed.

Condition 1 is itself an important security property, as shown for example in [21]. Condition 2 can be false depending on the context; for example, in a closed industrial control network the message set is usually limited; furthermore, messages in $\mathcal{M}$ can include integrity checks (as in the scheme discussed here) which alter the distribution of the plaintext. Condition 3 is another important security property, although some block ciphers have *weak keys* (e.g. DES and Blowfish). These conditions can be relaxed to having either the messages or the key uniformly distributed, provided that condition 1 still apply.

5.3.2 Computation of P_{re}

In this section, we assume these conditions are satisfied. Given that, as shown in section 3, the original message is $m = \mu\|\tau$, and the received message is $m' = \mu'\|\tau'$, the probability of residual error can be defined as the joint probability

$$P_{re} = P(\mu \neq \mu', \tau' = H(\mu')) \tag{7}$$

which correspond to the probability of having a correct integrity tag ($\tau' = H(\mu')$) but the decoded message is different from the original ($\mu \neq \mu'$) due to transmission errors. Considering that $ENC_{(k,m)}$ is a random map, each element of $\mathcal{C}'$ can correspond to all the possible values of $\mathcal{M}'$. If the elements of $\mathcal{C}'$ are uniformly distributed (such as when considering a random message attack) the probability that $\tau' = H(\mu')$, averaged over all the possible (k,m), can be computed by counting as

$$P^{avg}(\tau' = H(\mu')) = \frac{2^{\mu_{size}}}{2^{\mu_{size}+\tau_{size}}} = \frac{1}{2^{\tau_{size}}} \tag{8}$$

which is the probability of guessing a valid message. On the other hand, when transmission errors are considered, only one of the elements of $\mathcal{C}'$ correspond to the correct message, while all other elements have cumulative probability P_{re}^{ch}. The residual probability of error in this case can be computed as

$$P_{re}^{avg} = P_{re}^{ch}\frac{2^{\mu_{size}}-1}{2^{\mu_{size}+\tau_{size}}-1} \approx \frac{P_{re}^{ch}}{2^{\tau_{size}}}. \tag{9}$$

Both equations 8 and 9 are valid without making any assumption on the actual algorithm H used for computing τ.

This measure of P_{re} is, however, a measure for the average case, while from the safety point of view it is necessary to consider the worst case, that is:

$$P_{re}^{wc} = \max_{\substack{k\in\mathcal{K}\\ m\in\mathcal{M}}} P_{re} \tag{10}$$

This again can be computed by counting, but this time the actual distribution of $\mathcal{E}$, given by transmission errors, must be considered, specifically the one which maximise equation 10, which correspond to the worst-case pair (k^{wc}, m^{wc}). Using the ideal cipher model, the distribution in $\mathcal{E}$ can be obtained from a permutation of the distribution in $\mathcal{E}'$, so a simpler way to compute P_{re}^{wc} is to take the $2^{\mu_{size}}-1$ error vectors of $\mathcal{E}'$ with higher probability $P_{e'}$ and sum their probability. Clearly, the upper bound for the various $\mathcal{E}$ is

$$P_{re}^{wc} \leq P_{re}^{ch} \tag{11}$$

where equality would mean that the encryption procedure, even with an integrity check, is not effective at all in detecting transmission errors when (k^{wc}, m^{wc}) are used, because the relevant transmission errors e cause the plaintext to have an non-correctable error. More specifically, this correspond to the case where the $2^{\mu_{size}}-1$ most probable error vectors in $\mathcal{E}$ cover practically the whole amount of P_{re}^{ch}. The fact that this is independent from the particular algorithm H is because of the random map; in other words it is impossible to design an efficient integrity algorithm for this scheme as the distribution of e can not be known a priori.

This consideration exposes the trade-off between safety (as related to protection against random errors) and security (as related to protection against a malicious adversary). In the first case the statistical distribution of transmission errors is usually concentrated on a restricted set of values, and error detection codes are designed to detect the most probable errors, achieving a very high detection rate. On the other hand, if a malicious adversary is considered, all error patterns must be distributed ideally uniformly, otherwise an attacker may exploit its statistical characteristics.

5.3.3 Analytical results

To illustrate this problem, we consider a simplified model with a binomial distribution $\mathcal{E}' \sim B(n,k,p)$ (which would correspond in Figure 1 to the case with no channel coding and a Binary Symmetric Channel with probability of bit error p). The worst case P_{re}^{wc} can then be obtained by first listing the probability P_l

of each error vector e'_l with l bit set (note that must be $l > 0$); this list is then sorted incrementally and then the first values are taken, one error vector at a time, until exactly $2^{\mu_{size}} - 1$ error vectors are chosen. As shown in figure 2, where the worst-case residual probability of error is computed with the algorithm exposed above for different values of p in a binomial distribution, the value of P_{re} rapidly increases as the distribution of errors is more concentrated. Here τ_{size} is fixed to 64 bytes; similar results are obtained for different values. In figure 3 instead, μ_{size} is fixed to 256 bits and the worst-case P_{re} is plotted, and here the effect of the concentration of error distribution is also visible with effects similar to figure 2. This algorithm can be extended to the case where channel coding allows approximating the residual error distribution, for example if a CRC with a known minimum hamming distance H_{CRC} is used (some examples are available from the work of Koopman, see [22], for different payload lengths). In this case, the error vectors e_l with $l < H_{CRC}$ are not considered since they are detected by channel coding. However if channel coding include other information in the CRC (such as the length of the data payload) this approximation must be further refined.

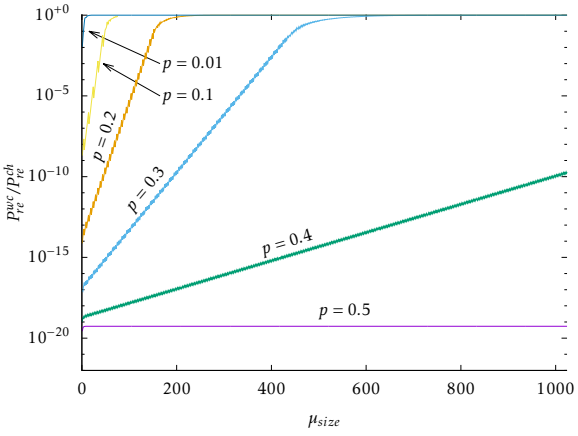


Figure 2: Normalised worst-case P_{re} obtained through computation, for different values of p. τ_{size} is fixed to 64 bytes.

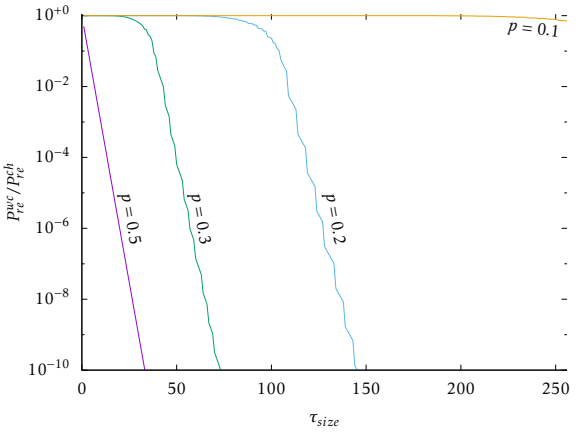


Figure 3: Normalised worst-case P_{re} obtained through computation, for different values of p. μ_{size} is fixed to 256 bytes.

5.3.4 Simulation results

The simulations have been performed applying the encryption scheme to randomly selected (k, m) and using different error vectors to obtain an approximation of P_{re}. For the ENC cipher, DES and AES have been used, while for H we used CRC, a truncated SHA1 hash function and a truncated HMAC scheme based on SHA1. The simulation have been implemented as a C++ program using the Nettle v2.7.1 cryptographic library. The simulations were run on a Intel i7 8-core laptop with 8 GB RAM; for each pair (k, m) a number of error patterns e' are generated, with the respective probability, and the normalised output P_{re} is computed as the ratio between the erroneous messages that resulted in $\tau = H(\mu)$ and the total number of erroneous messages. The plotted results represent a normalised $\frac{P_{re}}{P_{re}^{ch}}$, where the value of 1 represent equality in equation 11. The value of τ_{size} varies from 0 to 16; greater values, which in theory correspond to a lower P_{re}, have not been simulated since they would have taken too much time to yield a result with reasonable precision; however the results are still meaningful with respect to the theoretical model. The complexity of the simulation has three factors: k, m and e'. For example, using 1000 different keys, 1000 different messages and 10000 error patterns the total number of iterations is $1000 \cdot 1000 \cdot 10000 = 10^{10}$. However each value of P_{re} is evaluated using 10000 samples, so lower values close to 10^{-4} will have a lower accuracy. This explains the convenience to simulate with low τ_{size}.

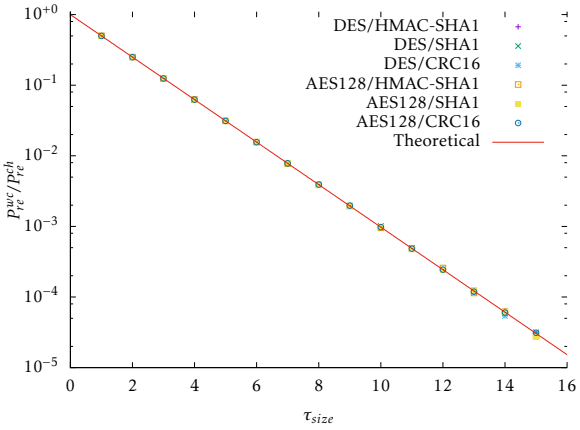


Figure 4: Normalised average P_{re} obtained through simulation and from equation 9, for different values of τ_{size}.

In Figure 4 the normalised P_{re}^{avg} is plotted, both resulting from equation 9 and from simulations. The correspondence between the theoretical model and the simulation results is very good, and the results are independent either from the H algorithm used to compute τ and the encryption algorithm ENC. Only a small glitch is visible for $\tau_{size} = 14$, presumably due to the relatively small number of iterations. The simulations results are accurate because all the P_{re} are av-

eraged over the pairs (k,m).

In Figure 5 and 6 the normalised P_{re}^{wc} is plotted with varying τ_{size}. The numerical values, displayed with a continuous line, are evaluated with the algorithm described in section 5.3.2, while the simulation results are taken as the highest value of P_{re} among all tested (k,m). In this case the simulations do not match the theoretical model; the reason is that while the theoretical model assumes that (k^{wc},m^{wc}) is known, in practice this is not true, although for some ciphers it could be feasible to calculate it. In this case a great number of (k,m) combinations are randomly chosen and the worst case is considered. However, being unable to scan all the (k,m) space, it is unlikely to find the worst case but only a "bad" pair (k,m) is found, for which P_{re} differ significantly from the average case.

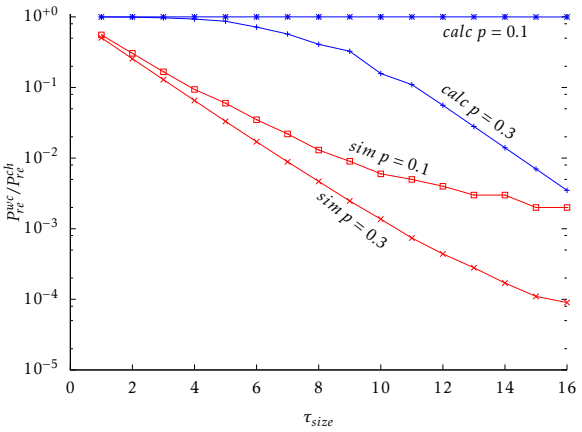


Figure 5: Normalised worst-case P_{re} obtained with $\mathcal{E}' \sim B(n,k,p)$ through calculation and simulation using a DES/SHA1 scheme, with $\mu_{size} + \tau_{size} = 64$, for different values of τ_{size}.

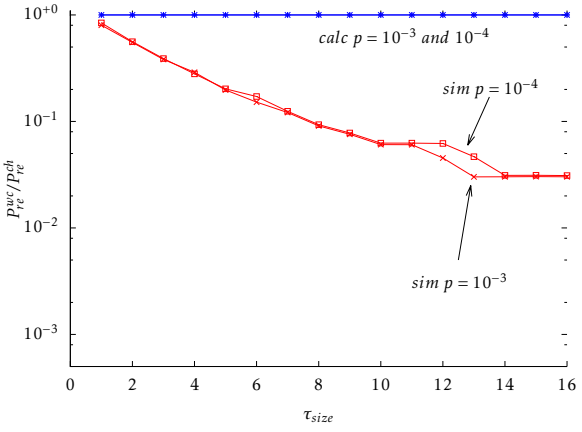


Figure 6: Normalised worst-case P_{re} obtained with $\mathcal{E}' \sim B(n,k,p)$ through calculation and simulation using a DES/SHA1 scheme, with $\mu_{size} + \tau_{size} = 64$, for different values of τ_{size}.

In Figure 7 and 8 the normalised P_{re}^{wc} is plotted with varying p. For low values of p, the value of P_{re}^{wc} does not change a lot, since the most probable error patterns are always the ones with 1 bit error. On the other hand, with higher p the value of P_{re}^{wc} approaches P_{re}^{avg}, which is reached with $p = 0.5$, corresponding to a uniform distribution. The issue of finding the worst case (k,m) is then different depending on the bit error probability of $B(n,k,p)$. For low p, approximately under 0.1, it is easier to find a pair (k,m) with high P_{re} since the most probable error patterns are the ones with only 1 bit error and are exactly $\mu_{size} + \tau_{size}$. On the other hand, for higher p, the most probable error patterns are a much great number, because it is easier to find more than one bit error. This explains the difficulty of finding the pair (k^{wc},m^{wc}) to simulate P_{re}^{wc} accurately.

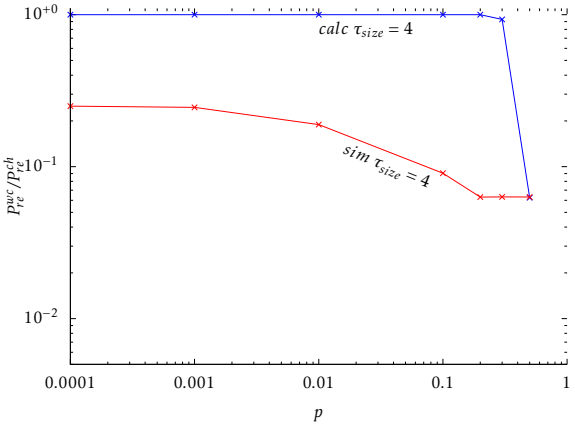


Figure 7: Normalised worst-case P_{re} obtained with $\mathcal{E}' \sim B(n,k,p)$ through calculation and simulation using a DES/SHA1 scheme, with $\mu_{size} + \tau_{size} = 64$, for different values of p.

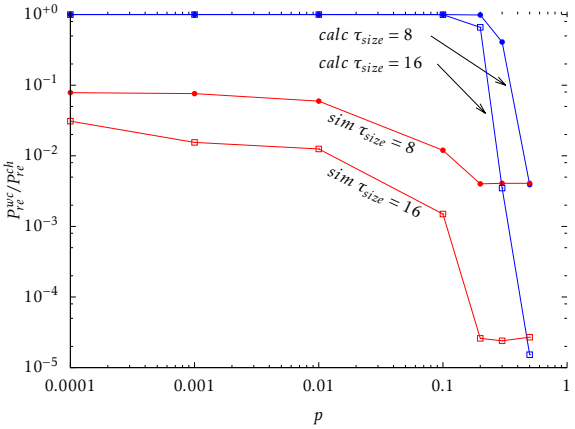


Figure 8: Normalised worst-case P_{re} obtained with $\mathcal{E}' \sim B(n,k,p)$ through calculation and simulation using a DES/SHA1 scheme, with $\mu_{size} + \tau_{size} = 64$, for different values of p.

6 Discussion

In Table 2 different message protection schemes are compared and ordered with decreasing security properties. The message protection schemes addressed

Scheme	Resists to	Security properties	Leaking out	Safety properties
ENC + MAC	Fully Malicious with chosen plaintext	Confidentiality Authentication Integrity	-	Strongly depends on (k, m)
ENC + CRC	Fully Malicious with chosen plaintext	Confidentiality Integrity	-	Strongly depends on (k, m)
plain + MAC	Fully Malicious	Authentication Integrity	Plaintext to FM and HBC attackers	Depends on $H()$
plain + CRC	Honest-But-Curious	-	Plaintext to FM and HBC attackers	Good under common channel assumption [13]
plain	-	-	Plaintext to FM and HBC attackers	-

Table 2: Summary of message protection schemes.

in this paper correspond to the ENC+MAC and ENC+CRC schemes, depending on the choice of $H()$, to resist a fully malicious attacker with chosen plaintext. The main alternative scheme, which does not consider the Confidentiality property, is evaluated for reference, based on literature work. Here the trade-off appears clear comparing the ENC+CRC and plain+CRC scheme; while the first has better security properties, the latter has better safety properties under common channel models, because CRC codes are designed specifically for correcting transmission errors.

7 Conclusion

In this paper, we have presented an analysis showing that a security property, like encryption, directly influences the probability of residual error, which is a safety property. On the other hand, the restricted size of the CAN bus payload forces the length of a MAC code to respect the fragmentation constraints which can be imposed by real-time requirements. With respect to similar schemes without encryption, by using a second CRC in addition to that one at the physical layer, the combination *ENC+CRC* performs worse; this is due to the intrinsic properties of block ciphers, which transform the distribution of errors to uniform on average. Other message protection schemes could have a less drastic impact on the worst-case error detection capability, or even this error detection capability could be embedded into the encryption algorithm itself, but then the risk is to offer the possibility for a side-channel attack.

Future works include the design and study of different message protection schemes, to offer a better trade-off between safety and security, for example improving the performance of the integrity tag τ with respect to transmission errors. Additionally, an experimental testbed on a real CAN bus would be useful to assess the performance of the protocol.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgement This work has been partially supported by the GAUSS (MIUR, PRIN 2015) and by the H2020 EU funded NeCS (GA 675320).

References

[1] G. Fortino, A. Rovella, W. Russo, C. Savaglio, "On the Classification of Cyberphysical Smart Objects in the Internet of Things" in *Proceedings of the 5th International Workshop on Networks of Cooperating Objects for Smart Cities (UBICITEC 2014)*, Berlin, Germany, Apr 14, 2014, pp. 86-94.

[2] G. Aloi, G. Caliciuri, G. Fortino, R. Gravina, P. Pace, W. Russo, C. Savaglio, "Enabling IoT interoperability through opportunistic smartphone-based mobile gateways" in *Journal of Network and Computer Applications*, Volume 81, 2017, Pages 74-84. https://doi.org/10.1016/j.jnca.2016.10.013

[3] C. Valasek and C. Miller, "Remote Exploitation of an Unaltered Passenger Vehicle" in *DEFCON 23*, 2015.

[4] A. Greenberg, "After jeep hack, chrysler recalls 1.4m vehicles for bug fix," Online, Jul 2015. [Online]. Available: https://www.wired.com/2015/07/jeep-hack-chrysler-recalls-1-4m-vehicles-bug-fix/

[5] B. McCluskey, "Connected cars: the security challenge for autonomous vehicles," *IET Engineering and Technology Magazine*, Feb 2017. [Online]. Available: https://eandt.theiet.org/content/articles/2017/02/connected-cars-the-security-challenge-for-autonomous-vehicles/

[6] C. Kim, "Safety challenges for connected cars," *IEEE Transportation Electrification Newsletter*, Jun 2016. [Online]. Available: http://tec.ieee.org/newsletter/june-2016/safety-challenges-for-connected-cars

[7] L. Dariz, M. Selvatici, M. Ruggeri, G. Costantino, and F. Martinelli, "Trade-off analysis of safety and security in CAN bus communication," in *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017, Naples, Italy, June 26-28, 2017*, pp. 226–231. [Online]. Available: https://doi.org/10.1109/MTITS.2017.8005670

[8] A. R. Chowdhury and S. DasBit, "Lmac: A lightweight message authentication code for wireless sensor network," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6. [Online]. Available: https://doi.org/10.1109/GLOCOM.2015.7417118

[9] N. Mouha, B. Mennink, A. Van Herrewege, D. Watanabe, B. Preneel, and I. Verbauwhede, *Chaskey: An Efficient MAC Algorithm for 32-bit Microcontrollers*. Cham: Springer International Publishing, 2014, pp. 306–323. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-13051-4_19

[10] N. Ferguson, "Authentication weaknesses in gcm," Comments submitted to NIST Modes of Operation Process, May 2005.

[11] Z. Gong, P. Hartel, S. Nikova, S.-H. Tang, and B. Zhu, "Tulp: A family of lightweight message authentication codes for body sensor networks," *Journal of Computer Science and Technology*, vol. 29, no. 1, pp. 53–68, 2014. [Online]. Available: http://dx.doi.org/10.1007/s11390-013-1411-8

[12] M. M. Fouda, Z. M. Fadlullah, N. Kato, R. Lu, and X. S. Shen, "A lightweight message authentication scheme for smart grid communications," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 675–685, Dec 2011. [Online]. Available: http://dx.doi.org/10.1109/TSG.2011.2160661

[13] F. Schiller and T. Mattes, "Analysis of nested crc with additional net data by means of stochastic automata for safety-critical communication," in *2008 IEEE International Workshop on Factory Communication Systems*, May 2008, pp. 295–304. [Online]. Available: http://dx.doi.org/10.1109/WFCS.2008.4638714

[14] A. Barenghi, L. Breveglieri, I. Koren, and D. Naccache, "Fault injection attacks on cryptographic devices: Theory, practice, and countermeasures," *Proceedings of the IEEE*, vol. 100, no. 11, pp. 3056–3076, Nov 2012. [Online]. Available: http://dx.doi.org/10.1109/JPROC.2012.2188769

[15] A. Rimoldi, "On algebraic and statistical properties of aes-like ciphers," Ph.D. dissertation, University of Trento, 2010. [Online]. Available: http://eprints-phd.biblio.unitn.it/151/

[16] S. Vaudenay, "Security flaws induced by cbc padding - applications to ssl, ipsec, wtls ..." in *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques: Advances in Cryptology*, ser. EUROCRYPT '02. London, UK, UK: Springer-Verlag, 2002, pp. 534–546. [Online]. Available: http://dl.acm.org/citation.cfm?id=647087.715705

[17] M. Vanhoef and F. Piessens, "Key reinstallation attacks: Forcing nonce reuse in WPA2," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM Press, Nov. 2017, pp. 1313–1328. [Online]. Available: http://dx.doi.org/10.1145/3133956.3134027

[18] M. Dworkin, "Recommendation for block cipher modes of operation: The cmac mode for authentication," NIST Special Publication 800-38B, May 2005.

[19] ISO, "Earth-moving machinery - machine-control systems (mcs) using electronic components - performance criteria and tests for functional safety," The International Organization for Standardization, Genève, Switzerland, Tech. Rep. ISO 15998, 2008.

[20] C. Petit, F.-X. Standaert, O. Pereira, T. G. Malkin, and M. Yung, "A block cipher based pseudo random number generator secure against side-channel key recovery," in *Proceedings of the 2008 ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '08. New York, NY, USA: ACM, 2008, pp. 56–65. [Online]. Available: http://doi.acm.org/10.1145/1368310.1368322

[21] L. Grassi, "New approaches for distinguishers and attacks on round-reduced aes," Cryptology ePrint Archive, Report 2017/832, 2017, http://eprint.iacr.org/2017/832.

[22] P. Koopman, "Best crc polynomials," 2017, https://users.ece.cmu.edu/~koopman/crc/.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 394-403 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Decentralized Control Approaches of Large-Scale Interconnected Systems

Rabeb Ben Amor*, Salwa Elloumi

Advanced Systems Laboratory in Polytechnic School of Tunisia, BP. 743, 2078, Marsa, Tunisia

ARTICLEINFO

Article history:
Received: 29 November, 2017
Accepted: 23 January, 2018
Online: 10 February, 2018

Keywords:
Interconnected systems
Decentralized stabilization
LMI

ABSTRACT

In this paper, we investigate the decentralized control problem for large-scale interconnected systems. The synthesis of the decentralized controller consists in determining gains which ensure the stability of the global system. To calculate these gains, three approaches are presented. Our main contribution is to develop a new decentralized stabilization approach which the decentralized local gains are calculated and formulated via the resolution of linear matrix inequalities (LMIs) problem. A numerical simulation comparison of the three methods is performed on an interconnected double-parallel inverted pendulum.

1 Introduction

This paper is an extension of the work originally we presented in the International Conference on Advanced Systems and Electric Technologies, 2017 [1]. This work treats three approaches dealing with the decentralized control of interconnected systems.

In fact, large-scale interconnected systems have received considerable attention in recent years due to its presence in several fields such as power electronics, robotics, communication, aerospace, transportation networks, manufacturing processes, biochemical applications and others. Designing a centralized control for these systems may not be efficient due to the modular nature of the system that can prevent the sharing of information between the various subsystems. Thus it is important to decompose the large-scale system into several subsystems. This decomposition which can be physical or mathematical, can make structures easier to control. This includes the implementation of decentralized control law.

In this way, it is necessary to decompose the global system into a number of interconnected subsystems for which, instead of a single centralized controller, a set of independent decentralized controllers is built. Thanks to its structure, the decentralized control has several advantages, mainly: the minimization of the information rate processed by the control units, the simplicity of the developed control laws compared to the centralized case and the improvement of the reliability of data transfer using only local information.

Many works in literature have been devoted to the decentralized control problems for interconnected systems. The decentralized adaptive control has been studied in [2–6]. The robust decentralized control is presented in [7–9]. The decentralized control using sliding mode approach is developed in [10–13].

Decentralized stabilization problem is the subject of our work. This problem is extensively studied in the literature and different design approaches were proposed accordingly [14–18]. To ensure the stability of the interconnected system formed by n subsystems, it is necessary to verify the local stability at each subsystem as well as the overall stability taking into account their interconnection.

The main contribution of this paper consists in developing some conditions allowing the synthesis of decentralized control laws that will ensure the stability of the overall interconnected system. In this way, we propose in this work a new decentralized stabilizing control approach for the interconnected systems. Indeed, the outcomes of this development are formulated in terms of linear matrix inequalities (LMIs).

The presented methods in our paper are applied to the physical system of two inverted pendulums interconnected by a spring. For the design of the decentralized control scheme, each pendulum should be seen as a subsystem. Many works used the typical system easily isolated into two subsystems to approve the validity of their proposed decentralized control approaches [19–22].

The rest of the manuscript is structured according to the following outline : The second section is reserved to formulate the problem and present the

*Corresponding Author : Rabeb Ben Amor, Email: rabebbenamor1@gmail.com, Salwa Elloumi, Email: salwa.elloumi@laposte.net

https://dx.doi.org/10.25046/aj030148

studied interconnected system formed by two parallel inverted pendulums coupled by a spring. In section 3, decentralized control approaches for the interconnected systems are presented, which are the decentralized quadratic optimal control and the decentralized pole-placement control. The last part of this section focuses on the development of a new decentralized stabilization control approach by using the Linear Matrix Inequalities Formulation. Section 4 is devoted to the implementation of the decentralized control approaches presented and developed in the previous section on the studied system. A comparative study between the three control approaches is presented to prove the validity of the new proposed approach. Finally, conclusions and some perspectives are given in the fifth section.

2 Problem Formulation and Description of the Studied Dynamic System

2.1 Problem Formulation

Large-scale interconnected systems are represented as follows:

$$\dot{x}_i = A_i x_i + B_i u_i + \sum_{\substack{j=1 \\ j\neq i}}^{n} H_{ij} x_j, i = 1, 2, .., n \qquad (1)$$

where $x_i \in \mathbb{R}^{n_i}$ and $u_i \in \mathbb{R}^{n_i}$ denote the state vector and the control vector of i^{th} subsystem, respectively.
$A_i \in \mathbb{R}^{n_i \times n_i}$ is the state matrix and $B_i \in \mathbb{R}^{n_i \times m_i}$ is the control matrix of each subsystem.
H_{ij} represents the term of interconnection between the i^{th} subsystem and the other subsystems.

The global interconnected system composed of N subsystems can be rewritten in a compact form as follows:

$$\dot{x} = Ax + Bu + Hx \qquad (2)$$

where:

- $x^T = [x_1^T, x_2^T, .., x_n^T]$ is the state vector of the global system ;
- $u^T = [u_1^T, u_2^T, .., u_n^T]$ is the control vector of the global system ;
- $A = diag[A_i], B = diag[B_i]$;
- H is the matrix formed by the terms of interconnection having the following form

$$H = \begin{bmatrix} 0 & H_{12} & \cdots & H_{1n} \\ H_{21} & 0 & & H_{2n} \\ \vdots & \vdots & \ddots & \\ H_{n1} & \cdots & \cdots & 0 \end{bmatrix}$$

2.2 Description of the Studied System : Double Inverted Pendulums Coupled by a Spring

We present in this section the description of the studied system formed by two interconnected inverted pendulum and its dynamic modeling.
In this system, two identical inverted pendulums of mass m directly mounted on the motor shafts in parallel where τ_1 and τ_2 are the input torques of each motor. These pendulums are connected to each other by an elastic spring of constant k which is mounted at the height a.
θ_1 and θ_2 are the angular displacements and of the pendulums from vertical.
New particular movements appear compared to the single movement of the individual pendulum. The interconnected inverted pendulums system is shown in figure 1 [23].

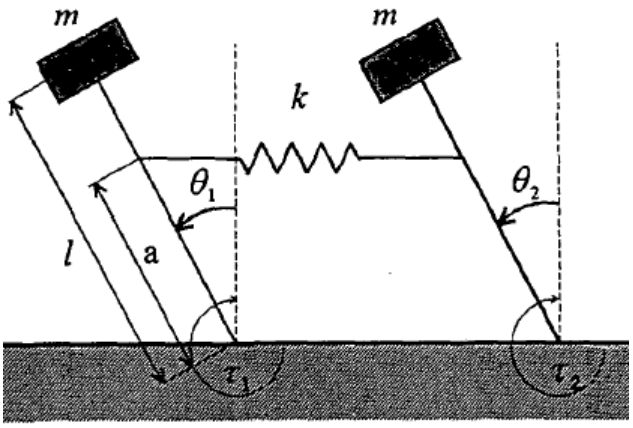


Figure 1: Modeling of the parallel inverted pendulum

The Lagrangian is defined as the difference between the kinetic energies and the potential energies of the system.

The kinetic energy for each pendulum is described by the following form:

$$T_i = \frac{1}{2} J_i \dot{\theta}_i^2 \qquad (3)$$

where J_i is the moment of inertia of the i^{th} pendulum and $\dot{\theta}_i$ is the angular velocity of i^{th} pendulum.

The total kinetic energy of the global system is then:

$$T = \frac{1}{2} J_1 \dot{\theta}_1^2 - \frac{1}{2} J_2 \dot{\theta}_2^2 = -\frac{1}{2} ml^2 \dot{\theta}_1^2 - \frac{1}{2} ml^2 \dot{\theta}_2^2 \qquad (4)$$

The potential energy for each mass is represented as follows:

$$V_i = mgl(1 - cos\theta_i) \qquad (5)$$

The potential energy of the spring is calculated using Hooke's law:

$$V_{spring} = \frac{1}{2}kx^2 = \frac{1}{2}k(-a\,sin\theta_1 + a\,sin\theta_2)^2 \quad (6)$$

The total potential energy of the system is given by:

$$\begin{aligned} V &= mgl(1-cos\theta_1) + mgl(1-cos\theta_2) \\ &+\tfrac{1}{2}k(-a\,sin\theta_1 + a\,sin\theta_2)^2 \end{aligned} \quad (7)$$

The Lagrangian of the interconnected studied system is written as follow:

$$\begin{aligned} L &= T - V \\ &= -\tfrac{1}{2}ml^2\dot{\theta}_1^2 - \tfrac{1}{2}ml^2\dot{\theta}_2^2 - mgl(1-cos\theta_1) \\ &-mgl(1-cos\theta_2) + \tfrac{1}{2}k(a\,sin\theta_1 - a\,sin\theta_2)^2 \end{aligned} \quad (8)$$

The Euler-Lagrange equations are given by:

$$\begin{cases} \frac{d}{dt}[\frac{\partial L}{\partial \dot{\theta}_1}] - \frac{\partial L}{\partial \theta_1} = \tau_1 \\ \\ \frac{d}{dt}[\frac{\partial L}{\partial \dot{\theta}_2}] - \frac{\partial L}{\partial \theta_2} = \tau_2 \end{cases} \quad (9)$$

Using Lagrange equations (9), we can easily show that the nonlinear equations of motion of the parallel inverted pendulum system are:

$$\begin{cases} -ml^2\ddot{\theta}_1 + mgl\,sin\theta_1 - ka^2[cos\theta_1(sin\theta_1 - sin\theta_2] = \tau_1 \\ -ml^2\ddot{\theta}_2 + mgl\,sin\theta_2 - ka^2[cos\theta_1(sin\theta_2 - sin\theta_1] = \tau_2 \end{cases} \quad (10)$$

Assuming a small angular displacement, the nonlinear equations of motion (10) can be replaced by the following linear model around the equilibrium point $\theta_1 = \theta_2 = 0$:

$$\begin{cases} -ml^2\ddot{\theta}_1 + mgl\,\theta_1 - ka^2(\theta_1 - \theta_2) = \tau_1 \\ -ml^2\ddot{\theta}_2 + mgl\,\theta_2 - ka^2(\theta_2 - \theta_1) = \tau_2 \end{cases} \quad (11)$$

So, the dynamics of the studied system composed of the two interconnected inverted pendulums are described by the following equations:

$$\begin{cases} -ml^2\ddot{\theta}_1 = mgl\theta_1 - ka^2(\theta_1 - \theta_2) - \tau_1 \\ -ml^2\ddot{\theta}_2 = mgl\theta_2 - ka^2(\theta_2 - \theta_1) - \tau_2 \end{cases} \quad (12)$$

For the design of the decentralized control scheme, each pendulum should be seen as a subsystem. Equations (12) can be writen into state equations with a standard choice of state variable for the i^{th} pendulum:

$$x_i(t) = \begin{bmatrix} \theta_i(t) \\ \dot{\theta}_i(t) \end{bmatrix}$$

The system consisted of two interconnected inverted pendulums is then described by the following state equations:

$$\begin{cases} \dot{x}_1 = A_1x_1 + B_1u_1 + H_1x_2 \\ \dot{x}_2 = A_2x_2 + B_2u_2 + H_2x_1 \end{cases} \quad (13)$$

with

- x_1, x_2 the state vectors of the subsystems ;
- u_1, u_2 the control vectors of the subsystem such as the input torque of each motor ;

The matrices and interconnection terms of the i^{th} subsystem are given by:

$$A_i = \begin{bmatrix} 0 & 1 \\ \frac{g}{l} - \frac{ka^2}{ml^2} & 0 \end{bmatrix}, B_i = \begin{bmatrix} 0 \\ \frac{-1}{ml^2} \end{bmatrix}$$
$$H_i = \begin{bmatrix} 0 & 0 \\ \frac{ka^2}{ml^2} & 0 \end{bmatrix}, i = 1, 2$$

The global system formed by two identical inverted pendulums coupled by a spring can be expressed by the following global state representation:

$$\dot{x} = Ax + Bu + Hx \quad (14)$$

where

- $x^T = [x_1^T, x_2^T]$ is the state vector,
- $u^T = [u_1^T, u_2^T]$ is the control vector.
- $A = diag(A_1, A_2)$ is the characteristic matrix:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{g}{l} - \frac{ka^2}{ml^2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{g}{l} - \frac{ka^2}{ml^2} & 0 \end{bmatrix}$$

- $B = diag(B_1, B_2)$ is the control matrix:

$$B = \begin{bmatrix} 0 & 0 \\ \frac{-1}{ml^2} & 0 \\ 0 & 0 \\ 0 & \frac{-1}{ml^2} \end{bmatrix}$$

- H is the term of interconnection:

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{ka^2}{ml^2} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{ka^2}{ml^2} & 0 & 0 & 0 \end{bmatrix}$$

with

m The mass of each pendulum, in Kg
l The rod length, in m
a The connecting position of the spring, in m
k The stiffness of spring, in N/m
g The acceleration of gravity, in $m.s^{-2}$
θ_i The angular displacement of the i^{th} pendulum, in Rad
τ_i The input torque of i^{th} motor, in $N.m$

3 Decentralized Control Approaches of Interconnected Systems

Possible control strategies for large-scale interconnected systems are generally based on a decentralized solution. A decentralized control structure applied to a process of n interconnected subsystems is shown in Fig 2.
The decentralized control partitions the measurement information and elaborates a local and independent control law for each subsystem.
It is necessary to check the stability of the interconnected system by examining two main aspects:
- Local stability: at each subsystem.
- Overall stability: taking into account the interconnections.

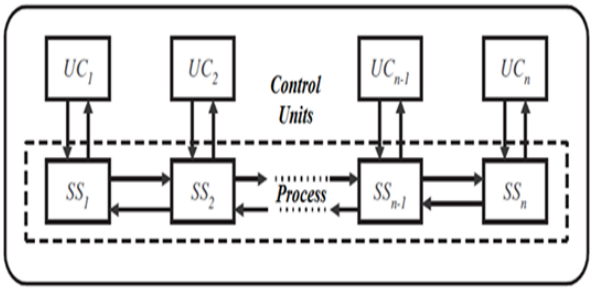


Figure 2: Decentralized control structure

The synthesis of the decentralized controller consists in determining the local gains K_i which ensure the stability of the overall closed-loop system.
To respect the decentralized information structure constraint, each subsystem is controlled by the local control law:

$$u_i(x) = -K_i x_i \quad i = 1,..,n \tag{15}$$

which leads to the following global control law of the overall system (2):

$$u(x) = -Kx \tag{16}$$

where $K = diag(K_1, K_2, .., K_n)$ is the block diagonal control gain matrix. Using global state-feedbacks, we get the closed loop system dynamics as following:

$$\dot{x} = A_f x \tag{17}$$

where:

$$A_f = \begin{bmatrix} A_1 - B_1K_1 & H_{12} & \cdots & H_{1n} \\ H_{21} & A_2 - B_2K_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & \cdots & \cdots & A_n - B_nK_n \end{bmatrix}$$

To calculate the gains K_i, different approaches can be considered.

3.1 Decentralized Quadratic Optimal Control

The decentralized control synthesis consists in considering the decoupled subsystems defined by the following state equations:

$$\dot{x}_i = A_i x_i + B_i u_i \tag{18}$$

and minimizing the modified quadratic criteria [24]:

$$J_i = \frac{1}{2}\int_0^\infty e^{2\alpha t}(x_i^T Q_i x_i + u_i^T R_i u_i)dt \tag{19}$$

Let $Q_i(n_i \times n_i), i = 1,..,n$ semi positive definite matrices, $R_i(m_i \times m_i), i = 1,..,n$ positive definite matrices and α a positive real.
The decentralized optimal control laws for each isolated subsystem can be expressed as a linear state feedback:

$$u_i = -K_i x_i, \quad i = 1, 2, .., n \tag{20}$$

where

$$K_i = R_i^{-1} B_i^T P_i \tag{21}$$

and P_i is the symmetric positive definite matrix solution of the following algebraic Riccati equation:

$$A_i^T P_i + P_i A_i - P_i(B_i R_i^{-1} B_i^T)P_i + 2\alpha P_i + Q_i = 0 \tag{22}$$

These decentralized state feedbacks applied to the interconnected system lead to the following global state representation:

$$\dot{x} = (A - BR^{-1}B^T P)x + Hx \tag{23}$$

where $R^{-1} = diag[R_i^{-1}]$ and $P = diag[P_i]$.
A sufficient condition to guarantee the stability of the overall system taking into account the interconnections, is given by the following theorem which proof is detailed in Appendix A.

Theorem 1 [24]:
The decentralized control law (16) is globally and asymptotically stabilizable for system (17) if the matrix F, given by(24), is positive definite.

$$\begin{aligned} F &= 2\alpha P + W - (PH + H^T P) \\ W &= Q + PBR^{-1}B^T P, \quad Q = diag[Q_i] \end{aligned} \tag{24}$$

3.2 Decentralized Pole-Placement Control

Pole-placement technique is a controller design method in which we determine the places of the closed loop system poles on the complex plane by setting a controller gain.
In this work we will apply this method for interconnected systems composed of n different subsystems that can be easily isolated. Firstly, it is necessary to verify the local stability.

For each subsystem, Ackermann's formula is used to find the control gain matrices.

Theorem 2: Ackermann's formula [25]
The controllability matrix C can be formed from:

$$C = [B\ AB...A^{n-1}B]$$

The feedback matrix K can be found as:

$$K = [0\ 0...1]C^{-1}P_d(A)$$

where P_d is the desired characteristic polynomial.

Using the local gain matrices obtained by Akermann's formula for each subsystem, the matrix in closed loop A_f of the global system taking into account the interconnection is given by:

$$A_f = \begin{bmatrix} A_1 - B_1K_1 & H_{12} & \cdots & H_{1n} \\ H_{21} & A_2 - B_2K_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & \cdots & \cdots & A_n - B_nK_n \end{bmatrix}$$

Stability condition:
In order to be stable, the eigenvalues of the system $\dot{x} = A_f x$ must all lie strictly in the left half of the complex s-plane. That means, the eigenvalues must all have strictly negative real parts.

3.3 Synthesis of a Decentralized Stabilization Control

This section deals with the global asymptotic stabilization of linear interconnected systems within the framework of Linear Matrix Inequalities (LMIs). We present the development of a new decentralized control approach.
To compute the gain matrix K, so that the closed loop system (17) is asymptotically stable, let consider the quadratic Lyapunov function represented by the following form:

$$V(x) = x^T Px \tag{25}$$

where P is a positive definite symmetric matrix of the following form:

$$A = \begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & P_n \end{bmatrix}$$

The time derivative of $V(x)$ is developped as :

$$\begin{aligned} \dot{V}(x) &= \dot{x}^T Px + x^T P\dot{x} \\ &= x^T A_f^T Px + x^T PA_f x \\ &= x^T (A_f^T P + PA_f)x \end{aligned} \tag{26}$$

The global asymptotic stability of system (17)provided with the decentralized control law (16) is ensured when the time derivative $\dot{V}(x)$ is negative definite which is equivalent to:

$$A_f^T P + PA_f < 0 \tag{27}$$

We note this expression by $\breve{A} = A_f^T P + PA_f$ with

$$A_f = \begin{bmatrix} A_1 - B_1K_1 & H_{12} & \cdots & H_{1n} \\ H_{21} & A_2 - B_2K_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & \cdots & \cdots & A_n - B_nK_n \end{bmatrix}$$

$\breve{A}$ can be written as:

$$\breve{A} = \begin{bmatrix} (A_1 - B_1K_1)^T & H_{12}^T & \cdots & H_{1n}^T \\ H_{21}^T & (A_2 - B_2K_2)^T & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1}^T & \cdots & \cdots & (A_n - B_nK_n)^T \end{bmatrix}$$

$$\begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & P_n \end{bmatrix} + \begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & P_n \end{bmatrix}$$

$$\begin{bmatrix} A_1 - B_1K_1 & H_{12} & \cdots & H_{1n} \\ H_{21} & A_2 - B_2K_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & \cdots & \cdots & A_n - B_nK_n \end{bmatrix}$$

So,

$$\breve{A} = \begin{bmatrix} A_1^T P_1 + P_1 A_1 - K_1^T B_1^T P_1 - P_1 B_1 K_1 & P_1 H_{12} + H_{21}^T P_2 \\ P_2 H_{21} + H_{12}^T P_1 & \ddots \\ \vdots & \vdots \\ P_n H_{n1} + H_{1n}^T P_1 & P_n H_{n2} + H_{2n}^T P_2 \end{bmatrix}$$

$$\begin{bmatrix} \cdots & P_1 H_{1n} + H_{n1}^T P_n \\ \ddots & \vdots \\ \ddots & \vdots \\ \cdots & A_n^T P_n + P_n A_n - K_n^T B_n^T P_n - P_n B_n K_n \end{bmatrix}$$

$$< 0 \tag{28}$$

Multiplying (28) on the right and then on the left by P^{-1} where P^{-1} is also symmetric positive definite matrix, inequality (28) becomes:

$$\begin{bmatrix} P_1^{-1}A_1^T + A_1P_1^{-1} - P_1^{-1}K_1^T B_1^T - B_1K_1P_1^{-1} & H_{12}P_2^{-1} + P_1^{-1}H_{21}^T \\ H_{21}P_1^{-1} + P_2^{-1}H_{12}^T & \ddots \\ \vdots & \vdots \\ H_{n1}P_1^{-1} + P_n^{-1}H_{1n}^T & \cdots \end{bmatrix}$$

$$\begin{bmatrix} \cdots & H_{1n}P_n^{-1} + P_n^{-1}H_{n1}^T \\ \ddots & \vdots \\ \ddots & \vdots \\ \cdots & P_n^{-1}A_n^T + A_nP_n^{-1} - P_n^{-1}K_n^T B_n^T - B_nK_nP_n^{-1} \end{bmatrix}$$

$$< 0 \tag{29}$$

It should be noted that the inequality matrix (29) has nonlinearities that are difficult to solve. We then use the changes of variables (30) and (31).

$$S_i = P_i^{-1} \tag{30}$$

$$L_i = K_iP_i^{-1} \tag{31}$$

Thus, the problem (29) can be rewritten to the form of linear matrix inequalities :

$$\begin{bmatrix} \breve{a}_{11} & \breve{a}_{12} & \cdots & \breve{a}_{1n} \\ \breve{a}_{21} & \breve{a}_{22} & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \breve{a}_{n1} & \breve{a}_{n2} & \cdots & \breve{a}_{nn} \end{bmatrix} < 0 \tag{32}$$

where:
$\breve{a}_{11} = S_1 A_1^T + A_1 S_1 - B_1 L_1 - L_1^T B_1^T$
$\breve{a}_{12} = H_{12} S_2 + S_1 H_{21}^T$
$\breve{a}_{1n} = H_{1n} S_n + S_1 H_{n1}^T$
$\breve{a}_{21} = H_{21} S_1 + S_2 H_{12}^T$
$\breve{a}_{22} = S_2 A_2^T + A_2 S_2 - B_2 L_2 - L_2^T B_2^T$
$\breve{a}_{n1} = H_{n1} S_1 + S_n H_{1n}^T$
$\breve{a}_{n2} = H_{n2} S_2 + S_n H_{2n}^T$
$\breve{a}_{nn} = S_n A_n^T + A_n S_n - B_n L_n - L_n^T B_n^T$

In order to find the gain matrices K of the decentralized control law, we have to solve the following LMI problem:

$$\begin{cases} S_i > 0 \quad i = 1, .., n \\ (32) \end{cases} \tag{33}$$

The following result is proved::
The interconnected system (17) provided with the decentralized control law (16) is asymptotically stable if LMI problem (33) is feasible.

4 Simulation results

This section is devoted to the implementation of the three decentralized control approaches exposed and developed in the previous section.
It consists in studying the stability by decentralized quadratic optimal control, decentralized pole-placement control and decentralized stabilization control based on LMI applied to the interconnected inverted pendulums system (Figure1), presented in Section2. The parameters of the studied system are summarized in Table 1.
In last party of this section, we carry out a comparative study between these three studied decentralized approaches to confirm the validity and the efficiency of the proposed approach.

Parameter	Value	Unit
m	0.4489	Kg
l	0.325	m
a	0.21	m
k	340.22	N/m

Table 1: The studied system parameters

Using the numerical parameters, model (14) of interconnected system composed of two parallel inverted pendulums is given by:

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -286.2486 & 0 & 316.4332 & 0 \\ 0 & 0 & 0 & 1 \\ 316.4332 & 0 & -286.2486 & 0 \end{bmatrix} x + \begin{bmatrix} 0 & 0 \\ -21.0903 & 0 \\ 0 & 0 \\ 0 & -21.0903 \end{bmatrix} u \tag{34}$$

To improve the performance of the studied system, we will apply the different studied approaches to guarantee an adequate stabilization.

4.1 Application of the decentralized optimal control approach

For this decentralized control, we focus on minimizing the modified quadratic criteria (19) for each separate pendulum.
The weighting factors are selected as follows:

$$\alpha = 0.2,$$
$$R_1 = R_2 = 0.0043$$
$$Q_1 = Q_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

The positive definite solution P_i of the Ricatti equation for each inverted pendulum is obtained by solving the equations (22):

$$P_1 = \begin{bmatrix} 0.0724 & 0.0014 \\ 0.0014 & 0.0002 \end{bmatrix}$$
$$P_2 = \begin{bmatrix} 0.0725 & 0.0014 \\ 0.0014 & 0.0002 \end{bmatrix} \tag{35}$$

Using (20) and (21) we obtain the decentralized control gain matrices:

$$K1 = [-7.0068 \ -0.8247]$$
$$K2 = [-7.0068 \ -0.8247]$$

To guarante the stability of the overall interconnected double-inverted pendulum, we should verify the theorem(24) when calculating the matrix F:

$$F = \begin{bmatrix} 0.3360 & -0.0278 & 0 & 0 \\ -0.0278 & 0.0030 & 0 & 0 \\ 0 & 0 & 0.3360 & -0.0278 \\ 0 & 0 & -0.0278 & 0.0030 \end{bmatrix}$$

The eigenvalues of the matrix F are given by:

$$\begin{bmatrix} 0.3383 \\ 0.0007 \\ 0.3388 \\ 0.0007 \end{bmatrix}$$

We can easily verify that matrix F is positive definite, so the decentralized control law stabilizes asymptotically the overall interconnected system (17).

The performances of the controlled system are shown in Figure3. The curves present the evolution of

the system state variables with decentralized optimal control, when some perturbations occur on θ_1 and θ_2. From the simulation results shown in these curves, it can be seen that the decentralized optimal control is able to enhance stability of the studied system in approximately 0.6 seconds.

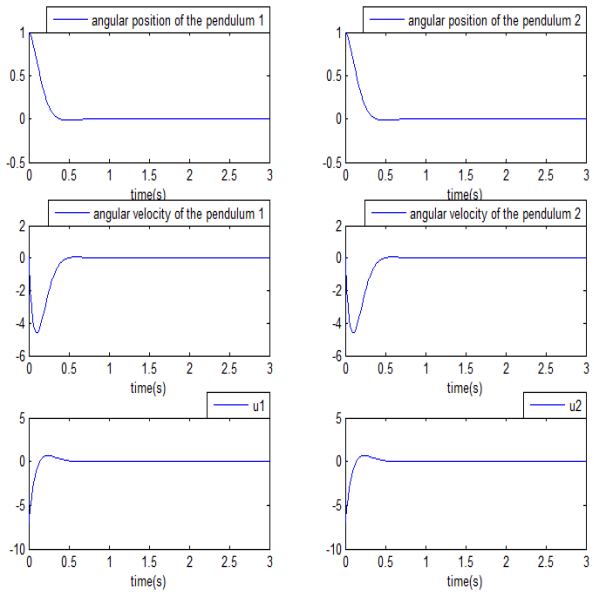


Figure 3: Evolution of the system state variables and corresponding decentralized optimal control signals

4.2 Application of the decentralized pole-placement control approach

In order to apply the decentralized pole-placement for the studied interconnected system, we shall firstly decompose the system into two decoupled inverted pendulums.
Thus, the dynamical model of the isolated subsystems is given by:

$$\dot{x}_1 = \begin{bmatrix} 0 & 1 \\ -286.24 & 0 \end{bmatrix} x_1 + \begin{bmatrix} 0 \\ -21.09 \end{bmatrix} u_1 \qquad (36)$$

$$\dot{x}_2 = \begin{bmatrix} 0 & 1 \\ -286.24 & 0 \end{bmatrix} x_2 + \begin{bmatrix} 0 \\ -21.09 \end{bmatrix} u_2 \qquad (37)$$

Since both (A_1, B_1) and (A_2, B_2) are controllable, we can move their poles to any desired locations, we choose the following desired eigenvalues:

$$\lambda_1^1 = -24; \lambda_2^1 = -18; \lambda_1^2 = -24; \lambda_2^2 = -12$$

In this case we calculate the local gains using the Ackermann's formula [25] :

$$K1 = [-6.9108 \ \ -1.9914]$$

$$K2 = [-0.0830 \ \ -1.7096]$$

Figures 4 and 5 present the evolution of the state variables and their corresponding pole-placement control signals for each isolated pendulum. From the simulation results shown in these curves, we can verify the local stability at each decoupled pendulum.

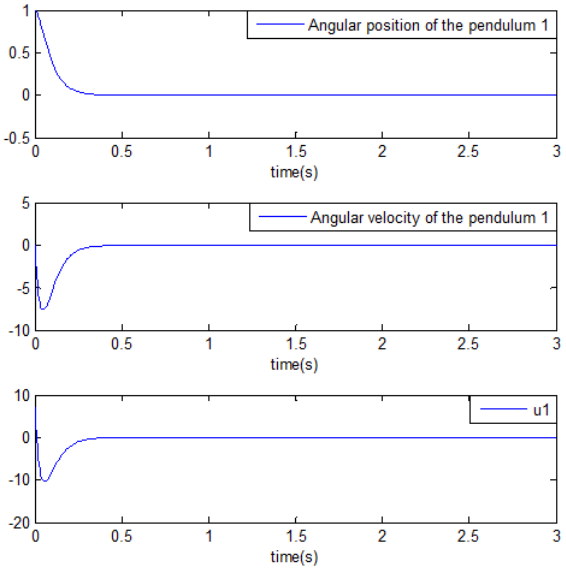


Figure 4: Evolution of the state variables and corresponding pole-placement control signals for the first decoupled pendulum

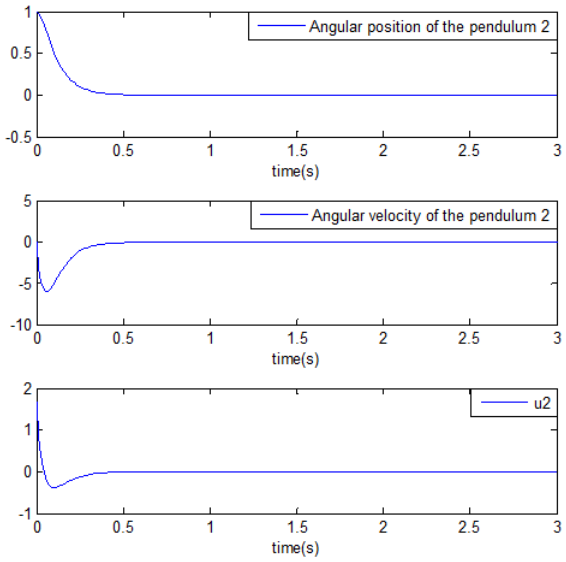


Figure 5: Evolution of the state variables and corresponding pole-placement control signals for the second isolated pendulum

After having applied the formula of Ackermann for each isolated decoupled pendulum, we obtain the closed loop matrix A_f of the overall interconnected system.

$$\dot{x} = A_f x = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -432 & -42 & 316.4332 & 0 \\ 0 & 0 & 0 & 1 \\ 316.4332 & 0 & -288 & -36 \end{bmatrix} x \qquad (38)$$

The eigenvalues of the system (38) have strictly negative real parts:

$$\lambda_1 = -0.9487$$
$$\lambda_{2,3} = -19.5639 \pm 17.0966i$$
$$\lambda_4 = -37.9236$$

Thus, the overall interconnected system provided with such a decentralized control law is asymptotically stable.
The curves in Figure6 illustrate the evolution of the system state variables and the corresponding decentralized pole placement control signals of the double inverted pendulum coupled by a spring, subjected to the same perturbations on the variable θ_1 and θ_2.
From the simulation results shown in these curves, it can be seen that the decentralized control is able to enhance stability of studied system in approximately 4 seconds.

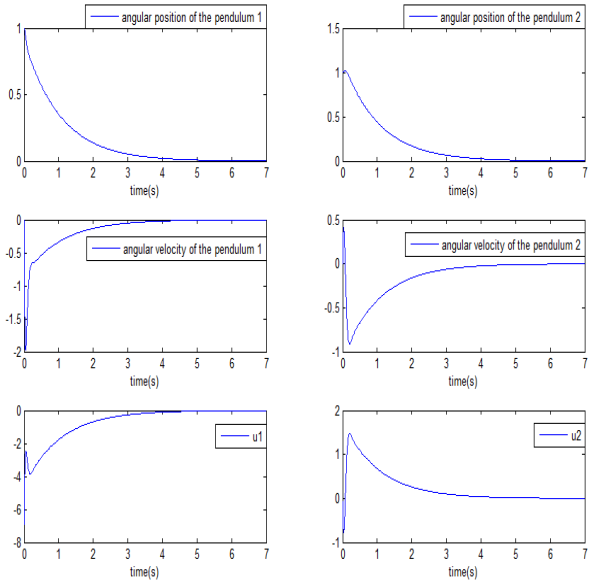


Figure 6: Evolution of the system state variables and corresponding decentralized pole-placement control signals

4.3 Application of the decentralized stabilization control approach

We consider the application of the proposed decentralized stabilizing control developed in section 3.3. on the studied system formed by two inverted pendulums coupled by a spring.
In this part, we solve the proposed LMI formulation in order to find the decentralized gains of the double inverted pendulum.
So we obtain:

$$\begin{cases} S_1 > 0 \\ S_2 > 0 \\ \begin{bmatrix} \breve{a}_{11} & \breve{a}_{12} \\ \breve{a}_{21} & \breve{a}_{22} \end{bmatrix} < 0 \end{cases} \quad (39)$$

where :
$\breve{a}_{11} = S_1 A_1^T + A_1 S_1 - B_1 L_1 - L_1^T B_1^T$
$\breve{a}_{12} = H_{12} S_2 + S_1 H_{21}^T$
$\breve{a}_{21} = H_{21} S_1 + S_2 H_{12}^T$
$\breve{a}_{22} = S_2 A_2^T + A_2 S_2 - B_2 L_2 - L_2^T B_2^T$

By solving problem LMI (39) we obtain the decentralized control gain matrices:

$$K1 = [-8.4386 \ \ -1.0841]$$

$$K2 = [-31.6377 \ \ -2.4752]$$

The evolution of the state variables of the dynamic system composed of two interconnected inverted pendulums with decentralized control by LMI is depicted in Figure 7.
It is clearly seen, from these curves, that the proposed decentralized stabilization control approach is efficient, it allows the best stabilization of the studied system despite the strong disturbances affecting the interconnection between its subsystems.

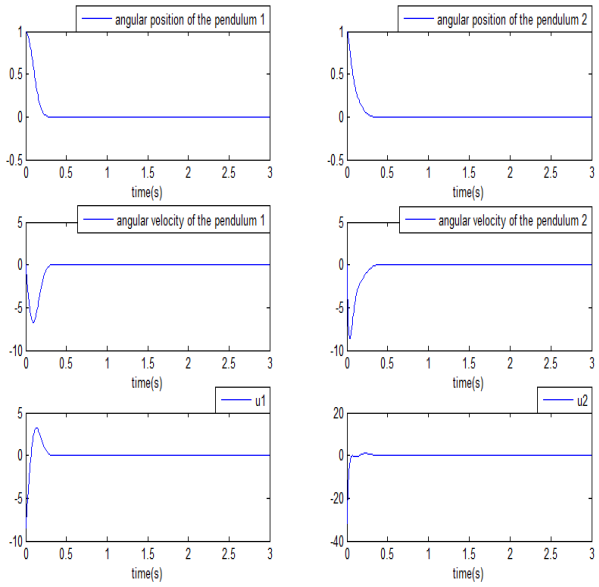


Figure 7: Evolution of the system state variables and corresponding decentralized stabilizing control signals

4.4 Comparative study of the three approaches

We present in this section a comparative study between the three decentralized control approaches studied in section 3.
The visualization of the curves in figure 3, 6 and 7 presenting the evolution of the system state variables and the corresponding control signals, submitted to the same perturbations, shows that the three studied decentralized control approaches can improve the stability of the interconnected system with double inverted pendulums coupled by a spring.

However, we find some disadvantages when applying the decentralized quadratic optimal and decentralized pole-placement on interconnected system. We were obliged to decompose the overall system into

a number of isolated subsystems and then to determine gain matrices that verify local stability for each subsystem. Then we present some sufficient conditions on the obtained gains to guarantee the global stability of the overall system taking into account the interconncetion terms. Indeed, we note an advantage for the proposed new stabilization decentralized approach using LMI problem which calculation of the local gains takes account of the interconnections terms.
On the other hand, when we compare the stabilization times of the three presented approaches, we note that our proposed approach is able to stabilize the system more quickly than other approaches.

5 Conclusion

This extended paper is devoted to the decentralized control techniques of large-scale interconnected systems. In this context, we have presented and studied some decentralized control approaches which objective is to synthesize the gains matrices in order to guarantee the stability of the global system. Our contribution focuses on the development of a new decentralized stabilization control approach based on linear matrix inequalities LMI.

The different approaches studied and formulated in this paper have been applied and validated on a double-parallel inverted pendulum coupled by a spring.
The simulation results have shown that it is possible to ensure the stability and improve the performance of the studied system controlled by each of the decentralized control laws relating to the proposed methods when some sufficient conditions are verified.
Comparative study presented in the fourth section has confirmed the validity and the efficiency of the proposed approach based on LMI which succeeded to ensure quickly the stability of the system and calculated the local gains taking account of the interconnections terms.

Many interesting directions for future research remain. One of the possible perspectives is to develop decentralized control nonlinear approaches for multi-robot cooperative system manipulating a common object.

Appendix A

The proof of the ***Theorem 1*** is based on Lyapunov direct method. Let V be the Lyapunov function defined by the following quadratic form:

$$V(x) = x^T P x \tag{40}$$

Using (23), The time derivative of V (x) is developped as :

$$\begin{aligned} \dot{V} &= x^T(A^T P - PBR^{-1}B^T P)x + x^T H^T P x \\ &\quad + x^T(PA - PBR^{-1}B^T)x + x^T PHx \end{aligned} \tag{41}$$

So (41) becomes

$$\begin{aligned} \dot{V} &= x^T(A^T P + PA - 2PBR^{-1}B^T P)x \\ &\quad + x^T H^T P x + x^T PHx \end{aligned} \tag{42}$$

Then, using the expression (22) in (42), we obtain:

$$\dot{V} = -x^T[2\alpha P + W - (PH + H^T P)]x \tag{43}$$

To ensure the asymptotic stability of system (23) , $\dot{V}$ should be negative definite, then which is equivalent to the matrix F:

$$\begin{aligned} &F = 2\alpha P + W - (PH + H^T P) \\ &W = Q + PBR^{-1}B^T P, \;\; Q = diag[Q_i] \end{aligned} \tag{44}$$

should be positive definite.

References

[1] R. Ben Amor, S. Elloumi, "On decentralized control techniques of interconnected systems-application to a double-parallel inverted pendulum" in International Conference on Advanced Systems and Electric Technologies(IC-ASET), Tunisia, 2017. https://doi.org/10.1109/ASET.2017.7983671

[2] C. Wang, C. Wen, Y. Lin, W. Wang, "Decentralized adaptive tracking control for a class of interconnected nonlinear systems with input quantization" Automatica 81 (2017) 359?368, 2017. https://doi.org/10.1016/j.automatica.2017.03.010

[3] R. Huang, J. Zhang, Z. Linc, "Decentralized adaptive controller design for large-scale power systems" Automatica 79 (2017) 93?100, 2017. https://doi.org/10.1016/j.automatica.2017.01.022

[4] H. Wu, "Decentralised adaptive robust control schemes of uncertain large-scale time-delay systems with multiple unknown dead-zone inputs" IET Control Theory Appl., Vol. 11 Iss. 9, pp. 1360-1370, 2017. https://doi.org/10.1049/iet-cta.2016.1277

[5] F. Sahami, M. U. Salamci, "Decentralized model reference adaptive control design for nonlinear systems; application to cancer treatment" in 17th International Carpathian Control Conference(ICCC), Slovakia, 2016. https://doi.org/10.1109/CarpathianCC.2016.7501173

[6] A. Hernndez-Mndez, J. Linares-Flores, H. Sira-Ramrez, "Decentralized adaptive control for interconnected boost converters based on backstepping approach" in Energy Conversion Congress and Exposition (ECCE), USA, 2016. https://doi.org/10.1109/ECCE.2016.7854996

[7] H. Wu, "Decentralised robust stabilisation of uncertain large-scale interconnected time-delay systems with unknown upper bounds of uncertainties" International Journal of Systems Science, 2015. https://doi.org/10.1080/00207721.2015.1029569

[8] H. Rtibi, S. Elloumi, "Robust decentralized nonlinear control for multimachine power systems" in International Conference on Advanced Systems and Electric Technologies(IC-ASET), Tunisia, 2017. https://doi.org/10.1109/ASET.2017.7983739

[9] H. Rtibi, S. Elloumi, N. Benhadj Braiek "A robust stabilization based on decentralized guaranteed cost control approach for a class of uncertain interconnected systems" in 4th International Conference on Systems and Control(ICSC), Tunisia, 2015. https://doi.org/10.1109/ICoSC.2015.7152782

[10] W. J. Liu, "Decentralized Sliding Mode Control for Multi-Input Complex Interconnected Systems Subject to non-smooth Nonlinearities" Asian Journal of Control, Vol. 20, No. 5, 2017. https://doi.org/10.1002/asjc.1627

[11] M. Fathallah , F. Abdelhedi, N. Derbel, "A synchronizing second order sliding mode control applied to decentralized time delayed multiagent robotic systems: Stability Proof" Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 160-170, 2017. https://doi.org/10.25046/aj020321

[12] F. Abdelhedi, N. Derbel, "Adelay-dependent distributed SMC for stabilization of a networked robotic system exposed to external disturbances" Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 513-519, 2017. https://doi.org/10.25046/aj020366

[13] M. Cucuzzella, G. P. Incremona, A. Ferrara, "Decentralized Sliding Mode Control of Islanded AC Microgrids with Arbitrary Topology" IEEE Transactions on Industrial Electronics Vol. 64, Iss. 8, 2017. https://doi.org/10.1109/TIE.2017.2694346

[14] Y. Li, J. Li "Decentralized stabilization of fractional order T-S fuzzy interconnected systems with multiple time delays" Journal of Intelligent and Fuzzy Systems 30 (2016) 319?331, 2016. https://doi.org/10.3233/IFS-151758

[15] J. Lin, "Decentralized Stabilization of Uncertain Large-scale Fractional Order Interconnected Systems via Output Feedback" in American Control Conference (ACC), USA, 2016. https://doi.org/10.1109/ACC.2016.7525336

[16] J. Liu, V. Gupta, "On Stabilization of Decentralized Systems Across Analog Erasure Links" IEEE Transactions on Automatic Control Vol. 62 Iss. 3, 2017. https://doi.org/10.1109/TAC.2016.2575833

[17] D. Liu , D. Wang, H. Li, "Decentralized Stabilization for a Class of Continuous-Time Nonlinear Interconnected Systems Using Online Learning Optimal Control Approach" IEEE Transactions on Neural Networks and Learning Systems, 2014. https://doi.org/10.1109/TNNLS.2013.2280013

[18] L. Bakule, M. de la Sen, M. Papik and B. Rehak, "Decentralized stabilization of symmetric systems with delayed observer-based feedback" in American Control Conference, USA, 2013. https://doi.org/10.1109/ACC.2013.6580888

[19] H. Ferdowsi, S. Jagannathan, "Decentralized Fault Tolerant Control of a Class of Nonlinear Interconnected Systems"International Journal of Control, Automation and Systems 15(1) (2017) 1-10, 2017. https://doi.org/10.1007/s12555-015-0384-5

[20] B. Zhao, D. Wang, G. Shi, D. Liu, Y. Li, "Decentralized control for large-scale nonlinear systems with unknown mismatched interconnections via policy iteration" IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017. https://doi.org/10.1109/TSMC.2017.2690665

[21] B. Zhao, D. Liu, Y.Li, Q. Wei, R. Song, "Adaptive Dynamic Programming Based Decentralized Tracking Control for Unknown Large-scale Systems" in Proceedings of the 36th Chinese Control Conference, China, 2017. https://doi.org/10.23919/ChiCC.2017.8027913

[22] C. Wei, J. Luo, H. Dai, Z. Yin, J. Yuan, 'Low-complexity differentiator-based decentralized fault-tolerant control of uncertain large-scale nonlinear systems with unknown dead zone"Nonlinear Dynamics , Vol. 89, Iss. 4, 2017. https://doi.org/10.1007/s11071-017-3605-z

[23] U.S. Park, J.H. Kim, S.J. Kim, J.W. Choi, J.S. Kim, "Development of a Parallel Inverted Pendulum System and Its Control" in 38th SICE Annual Conference, Japan, 1999. https://doi.org/10.1109/SICE.1999.788676

[24] M.S. Mahmoud, M.F. Hassan, M.G. Darwish, "Large scale control systems, theories and techniques" Edition Marcel Dekker, New York, 1985.

[25] J. Ackermann, "Sampled-Data Control Systems" Springer-Verlag, Germany, 1985.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 404-411 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Security Analysis and the Contribution of UPFC for Improving Voltage Stability

Asma Meddeb*, Hajer Jmii, Souad Chebbi

LATICE laboratory of National High School of Engineers of Tunis, Tunis 1008, Tunisia, Tunis University.

ARTICLE INFO

Article history:
Received: 30 October, 2017
Accepted: 02 January, 2018
Online: 10 February, 2018

Keywords:
Power system security assessment
Contingency analysis
Newton Raphson power flow method
UPFC

ABSTRACT

The occurrence of many failures in the power system can lead to power instability and affects the system parameters to go beyond its operating limits. It may lead to obstructing the secure operations and reliability of power systems. Ensuring power system security needs proper actions to be taken for the undesirable contingency. Thus, security analysis is important tasks in modern energy management systems. This paper proposes an approach based on the Newton Raphson power flow method for power system security analysis. Firstly, the contingencies will be specified to assess their impact on the transient stability. Secondly, the selected contingencies will be classified in the order of severity. In addition, the integration of the Unified Power Flow Controller (UPFC) to enhance the transient stability of the power system is considered. The proposed method is implemented on the IEEE-14 bus system. We performed this case study using the well-known software EUROSTAG.

1. Introduction

Security analysis of a power system plays a significant role in the growth and development of modern societies. Given that the power system is a large-dimension complex system, it might be the site of various disturbances, making its chaotic behavior and difficult to control [1]. Thus, ensuring the security of the electric network becomes an essential task for operators and researchers.

Following a violent disturbance enormous damage may occur in the power system and impacts directly its normal operation. We can note that the concept of network security is broader than the stability of networks [2]. It is defined as the ability of power system to maintain its stable operation after each disturbance or unfavorable change of the system [3]. However, the level of security is deemed satisfactory if, firstly, it is able to ensure the production-consumption balance within certain acceptable limits related to line's power flows, voltage in each node and operating points of generation units and if, on the other hand, it is able to persist in a more or less violent disturbance.

Therefore, the system operator must ensure that the power system operates at a secure operating point where all equipment is within their safe limits. If the failure is not controlled suitably, this could lead to a disastrous event such as major blackouts and impact then the economic and environmental requirements [4].

*Corresponding Author: Asma Meddeb, National High School of Engineers of Tunis (ENSIT), Tunis 1008, Tunisia, +21697467529, asma.meddeb@ensit.rnu.tn

The advent of FACTS devices has been a boon to the power industry. These devices provide fast and effective control of the various parameters of the power system to improve their stability and preventing other related cascade accidents [5-9]. The UPFC is one of the most versatile FACTS devices, providing independent control of the voltage, real and reactive power of the transmission under its supervision. This device enhances the power system security.

N. Hashim *et al*. [10] have presented an analysis of transient stability of IEEE 14 bus test system. They studied the effect of the fault location on system stability; and they analyzed the characteristics of the machine states including speed, rotor angle, output electrical power and the terminal voltage of the machine after a three-phase short-circuit which occurs at different points of the system. The IEEE-14 bus system has been also studied in [11], wherein both authors have studied the effect of fault location and critical clearing time on the system stability. In order to achieve this, they analyzed the behavior of the synchronous machine, in particular, the angular position of the rotor. They proposed to rapidly isolate the faulted part to increase stability margin.

A. Zerigui [12] provided a solution to the problem of transient stability. To test a constrained optimal power flow, and estimate critical clearing time she developed a new analytical function. She applied her strategy to test large networks up to 145 buses. While in [13] the authors just used nine-bus IEEE system to confirm their proposed method. They presented a global approach to transient stability constraint optimal power flow based on SIME model.

www.astesj.com
https://dx.doi.org/10.25046/aj030149

Likewise, authors in [14] analyzed the transient stability for IEEE-9 bus test system. Then, they proposed a simple control method based on the equal area criterion for calculating critical clearing time. To improve the stability margin, the temporal evolution of the frequency and voltage, with and without load shedding, is performed.

For three decades, J-C Chow et al [15] have formulated the contingency classification problem into a pattern recognition problem and then design a Hopfield model to detect a prescribed set of patterns. This optimization method, which uses the linear programming technique, considers only the steady-state security assessment problem.

In order to explain the contribution in deepening understanding of the power system security problem, the authors in [16], proposed an efficient steady-state contingency classification methodology based on Rough Set Theory. Through this analysis, they could classify the system operating in four different states.

Abdulrazzaq, A. [17] used the Newton Raphson load flow method for the power system contingency ranking only for the line outage.

In most cases, the researchers deal with the security analysis in the steady case, considering the execution time of computational simulations performed. The main purposes of security analysis are the fast identification of critical contingencies and their evaluation related to the severity level [18-20].

This paper is an extension of work originally published in the Proceedings of the International Conference on Sciences of Electronics, Technologies of Information and Telecommunications [21].

This paper focuses on the contingency analysis that comprises three steps. Firstly, we create the contingencies list containing all sets of possible contingencies that may occur in a power system. Secondly, the classification of severe contingencies from the list that may lead to the supply and security constraints violations is achieved. Lastly, the evaluation of contingencies is necessary. In addition, a curative action using UPFC devices to avoid the effects of most severe contingencies is proposed. To verify our proposed approach, we perform simulations on an IEEE 14-bus power system.

2. Analysis of Electrical Networks security

The security analysis of an electrical network is related to its dynamic state or steady state following the disturbance. In dynamic security analysis, the transition from the existing operating condition for the new operating conditions and the fact that during the transient state should not be cascading outages are considerations of interest. The dynamic security analysis is normally done in a deterministic approach using analytical tools such as load flow calculation, dynamic simulation, etc. Indeed, these tools use a detailed model of the electrical system to determine the system dynamic response with respect to each of the analyzed contingencies. Analysis methods of the dynamic security of electrical networks are classified into four groups: methods based on numerical simulation, pattern recognition methods, direct methods and hybrid methods [22-25]. In this paper, we propose heuristic method because it is easier to set up and it gives quick and good solutions to difficult problems.

2.1. Operating States of a Power System

In security analysis, power system may operate in different states, namely normal, emergency, alert, extreme state and restorative [26, 27]. A system is said to be in normal operating state when it satisfies equality and inequality sets of constraints. This is equivalent to the fact that there is no overloaded equipment and all the variables of the system are within the normal operating limits [28, 29]. Following every single contingency the system must be able to operate in a secure way while respecting all the constraints. However, if the contingency is highly dangerous or the system generation drops below the required amount, the system operation may be insecure indicating an alert state. In this case, the equality and inequality constraints are still satisfied nevertheless; preventive action is needed to restore the normal state. In case of preventive control failure, the system security level may be under the permissible limit. Therefore, the inequality constraints are no longer maintained and the system enters in an emergency state. It is to be mentioned that in case of a sufficiently serious contingency, the system transits directly to the extremis state.

To overcome the emergency state, control measures should be initiated immediately. If these actions are not efficient or are not taken at the adequate time, the state of the system is said in extremis. This state is characterized by violated equality and inequality constraints which may cause the loss of some parts of the system or a total system blackout. Otherwise, the rapid and efficient application of control actions brings the system to the restorative state. In that instance, the implementation of other control measures may lead the system either to the normal state or the alert state.

2.2. UPFC device

UPFC (Unified Power Flow Controller) is the most powerful and versatile device. It can control three parameters either individually or in appropriate combinations at its series connected output while maintaining reactive power support at its shunt connected input device. Therefore, the aim of UPFC is to enhance the usable transmission capacity of lines and control the power flow [30-35].

The model of UPFC implemented in a transmission line is shown in Figure 1. A series converter connects with the line through an insertion transformer. A shunt converter connects with the line through a second transformer. The DC terminals of the two converters are linked together, and their common DC voltage is supported by a capacitor bank. A mathematical model of the UPFC is well-detailed in Reference [36].

3. Mathematical Model of power system

In order to analyze, simulate, design and control the electric power system operation, steady-state and dynamic system performances properly are explained.

3.1. Description of system

In general, a power system can be modeled by the following set of nonlinear differential-algebraic equations:

$$\dot{x} = F(x, y, \lambda)$$
$$0 = G(x, y, \lambda) \quad (1)$$

Where x is a vector of state variables, y is a vector of network variables, and λ represent the control and parameter variable which may be used to control or tune power system performance.

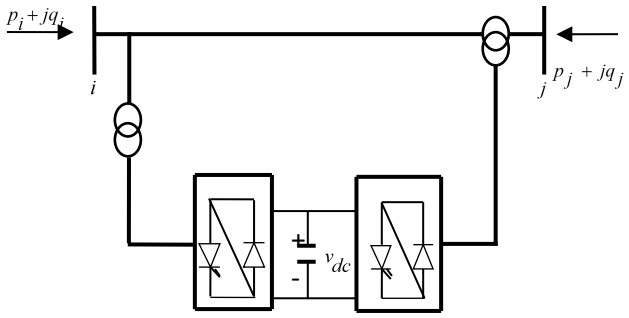


Figure 1. UPFC model.

The differential equation F represents governing dynamics of power systems, which is associated with generators, excitation systems, and speed governor. The differential variables x consist of the states of dynamic components such as rotor angles and rotor speed.

$$x = [\delta, \omega]^T \quad (2)$$

The algebraic equation g represents the network power balance of power systems. The algebraic states y includes bus voltages and bus phase angles. It is also known as control variables.

$$y = [V, \alpha]^T \quad (3)$$

The active and reactive powers injected at any bus are given by Equation (4). Therefore, it is a function of voltage magnitudes and angles for all the buses, as well as the admittance, magnitude, and angle, for the transmission network.

$$\begin{cases} P_i = V_i . \sum_{i=1}^{n} Y_{ij} . V_j . \cos(\alpha_i - \alpha_j - \theta_{ij}) \\ Q_i = V_i . \sum_{i=1}^{n} Y_{ij} . V_j . \sin(\alpha_i - \alpha_j - \theta_{ij}) \end{cases} \quad (4)$$

Where Y_{ij},θ_{ij} are, respectively, the admittance magnitude and angle for the transmission line i-j, V_i, V_j are the voltages at bus i and j, α_i, α_j are the voltages angles, and n is the number of studied network buses.

The active and reactive power functions are evaluated by this load flow equations:

$$\begin{cases} P_{Gi} - P_{ci} - P_i = 0 \\ Q_{Gi} - Q_{ci} - Q_i = 0 \end{cases} \quad (5)$$

P_{ci}, Q_{ci}, P_{Gi} and Q_{Gi} are respectively active and reactive power consumed and produced for bus i.

Equations (6) give the output active and reactive power. These equations show that P and Q are functions of the bus voltages and rotor angle. X_T and E_q represent respectively transfer reactance between buses and internal voltage of the generator.

$$\begin{cases} P = \frac{E_q . V_i}{X_T} . \sin\delta \\ Q = \frac{V_i^2}{X_T} - \frac{E_q . V_i}{X_T} . \cos\delta \end{cases} \quad (6)$$

3.2. Model of Machine

The generator model is represented by the four-order model as follows:

$$\begin{cases} \frac{d\delta}{dt} = \omega - 1 \\ M . \frac{d^2\delta}{d^2 t} + \frac{d\delta}{dt} . D = P_m - P_e \\ \frac{dE'_q}{dt} = \frac{1}{T'_d} . (E_{fd} - E'_q + (X_d - X'_d) . i_d) \\ \frac{dE'_d}{dt} = \frac{1}{T'_q} . (-E'_d + (X_q - X'_q) . i_q) \end{cases} \quad (7)$$

δ is the rotor angle of the machine, ω is the rotor speed, M is an inertia coefficient of machine, D represents the damping coefficient of machine, Pm, Pe are the mechanical and electrical power of the machine, E'_d, E'_q are the voltage behind the direct and quadrature axis transient reactance X_d', X_q' respectively, T'_{d0} is the d-axis open circuit transient time constant, T'_{q0} is the q-axis open circuit transient time constant, and E_{fd} represents a field voltage.

3.3. Security Contraints

The security constraints are based on the operating limits which have to be satisfied for normal operation of the power system. In order to operate the system within an acceptable security domain, the basic security constraints are as follows:

$$V_{min} \le V \le V_{max} \quad (8)$$

$$P_{gmin} \le P_g \le P_{gmax} \quad (9)$$

$$Q_{gmin} \le Q_g \le Q_{gmax} \quad (10)$$

$$\delta_{min} \le \delta \le \delta_{max} \quad (11)$$

4. Proposed Method

Considering the fact that the electrical network operates under two main types of constraints; constrained supply (C1) of energy (C2) which imposes that all consumers must be supplied and operations constraints require that the system variables should be within specified operating limits.

During a normal operating state, all constraints are satisfied. It is characterized by a sufficient level of stability margins so that the system can withstand a single contingency. When the system satisfies all the supply and security constraints without reserve power generation, it passes in an alert state. This means that in the event of a contingency, at least one inequality constraint will be violated. Therefore, preventive actions have to be taken to bring the system to a secure state. The power system enters an emergency state from an alert state when a contingency occurs. In this case, all the equality constraints are satisfied and at least one inequality constraint is violated. Consequently, the system requires an immediate implementation of corrective actions to prevent the damage of equipment. However, once the severity of the disturbance is very high, both equality and inequality constraints are violated. Thus, the system passes in an extreme state.

In this paper, we consider the time taken by the system to transit from a secure state into another one where is not secure as an index for ranking serious contingencies. Therefore, the contingency is deemed severe when both constraints of the network are not satisfied and, it is potentially harmful if only the constraints of system operation are not satisfied. In the case of both constraints are respected, the contingency is recognized as being harmless.

To achieve this aim, a fast strategy is proposed to identify power system dynamic behavior using the time domain simulation method. A corrective control to restore power system operating equilibrium after disturbances has been also proposed. This strategy is based on a deterministic approach using analytical tools such as load flow calculation and dynamic simulation. Newton-Raphson method is used to solve load flow problem. In this method, optimal solution can be obtained using iterative method. It usually converges faster than other methods. Then, the identification of properties by the best solutions and introduce them as boundaries of the problem are mainly fulfilled. Firstly, the assessment of different operating state according to a set of selected contingencies is required. Then, security assessment is based primarily on steady-state load flow analysis and transient analysis of the power system. The based approach gives priority to the most severe cases for thorough analysis, and the stable cases are previously eliminated from the list.

Figure 2 depicts the flowchart of the proposed method.

5. Simulation

5.1 Case Study

Figure 3 shows the single line diagram of IEEE 14-bus system. It consists of five synchronous machines. There are eleven loads in the system. Three step-up transformers, one of which is three winding transformer. The generators are modeled as an ideal voltage source behind the synchronous reactance of the machines. The model of transmission lines considers the resistance and the reactance. The loads are modeled as constant impedance. The all data for simulation were selected from [37]. We performed all simulations using software package EUROSTAG [38]. This software is a powerful tool dedicated to dynamic simulations. Its major advantage is the high rapidity of its algorithm.

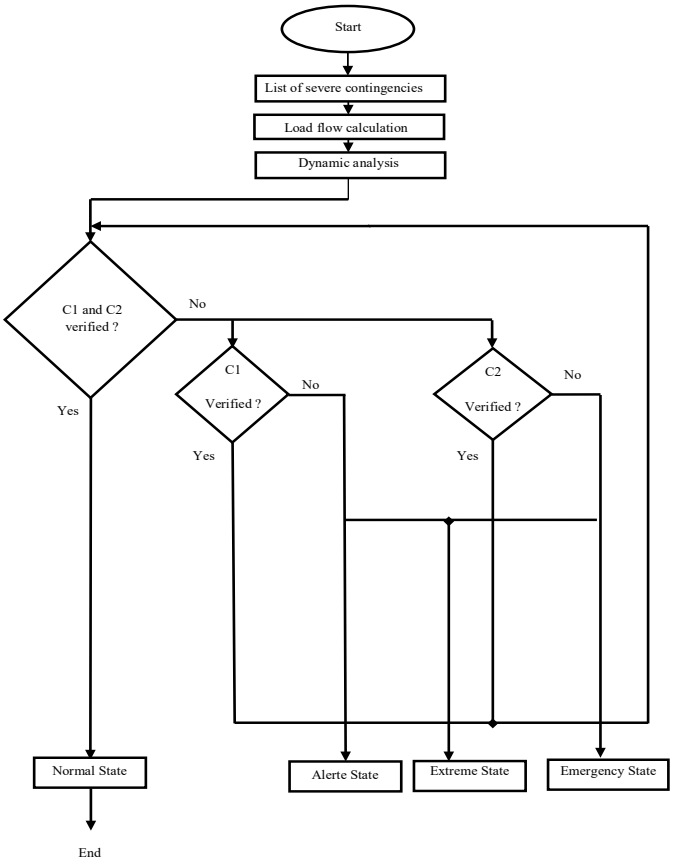


Figure 2. Flow chart of proposed method.

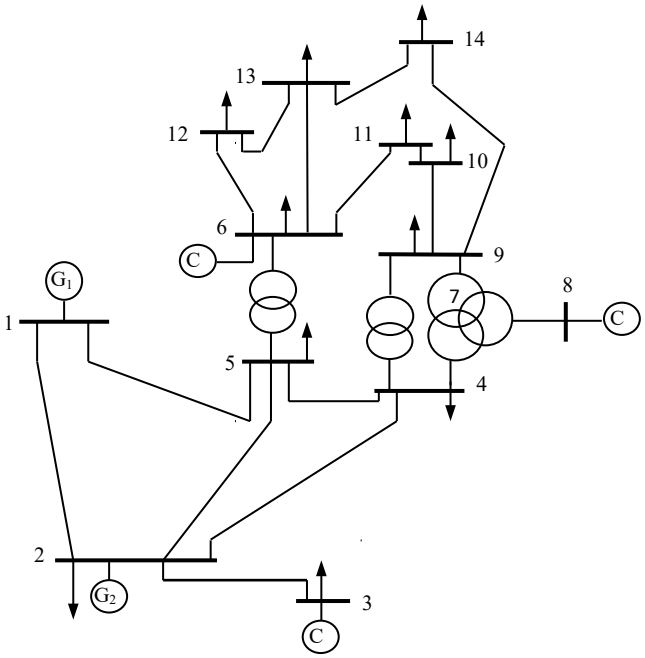


Figure 3. Single line diagram of IEEE 14-bus system.

5.2 Contingency Classification Method

Contingency analysis should be performed for the unexpected and severe events that may occur in power system. Table 1 presents the contingencies list containing the contingencies that may occur in a power system. For the contingencies classification, we considered two criteria; maintaining the network characteristics in their acceptable ranges, and ensuring a balance between production and consumption. A contingency is deemed very dangerous when both constraints of the network are not satisfied, which corresponds to an extreme state. It is potentially dangerous if only the constraints of system operation are not satisfied; the network is in an emergency state. Whenever the system goes to alert state under severe contingency; is that the constraint of energy supply is not fully respected. If both constraints are respected the contingency is recognized as being harmless; as a result, the network is stable. Thus, this contingency is ineffective on the loss of system and it can be rejected.

Table 1. Contingency Index.

Contingency	Index
Bolted fault	I
Impedant fault	II
Increase of load by 10%	III
loss of generator G1	IV
loss of generator G2	V
Opening of line 8-7	VI
Opening of line 2-3	VII

5.3 Simulations Results

At time 250s, we simulated different types of fault and we followed the temporal transition of the system. Table 2 defines the emergency state, alert and the extreme state of network study. We can notice that the range of stability is different for each type of fault. Loss of generator G1 remains the scariest of all faults on network stability. In order to have reasonable accuracy in the classification of contingencies, the detailed assessment of the contingencies is necessary. The aim is to detect the worst case for the transient stability of the test system. Thus, we tested the impact of line opening, loss of power generator, and the severity of three-phase short-circuits.

5.3.1 Line opening scenario

We should first look at how the line opening affects the transient stability of the system. To achieve that, we propose the tripping of lines one by one. All system parameters (frequency, voltages at different buses, rotor angle of the two machines) must be within the allowable range to predict the security of the system.

Therefore, all security practice is followed. For the opening of line 1-5, the behavior of the two machines is different; G1 accelerates while the machine G2 decelerates. Both machines accelerate, in case VII and oscillations are much less than line 1-5. In the case of opening the line 9-14, a voltage drop at bus 14 passes from 0.971pu to 0.928pu, oscillations amplitude of the rotor angle is important. With the opening of line 8-7, we note a separation of the network in two sub-networks; the first sub-network includes the machines G1, G2, G3, and G6. The second involves only the machine G8. Following this fault, a drop in voltage appears at the terminals of all the consumer nodes and an increase of 2.5% of the voltage at bus 8 with a slight oscillation of voltages.

Table 2. Operating modes of the test system

State	Alert State	Emergency state	Extreme State
I	[250.1 250.4]	] 250.4 250.413]	]250.413 250.414]
II	[251.5 252]	] 252 253]	] 253 255]
III	-	-	-
IV	] 250.07 250.114]	] 250.114 250.13]	] 250.13 250.15]
V	] 250.9 251.5]	] 251.5 257]	] 257 260]
VI	-	-	-
VII	-	-	-

We note that the variation in the speed and power of the two generators have a damped oscillatory regime. Then we can say that the system is stable under these conditions because generator speed and hence rotor angle regain their stability after few highly damped oscillations. In conclusion, the opening of a single line does not disrupt the stability of the network; this contingency may be classified as non-dangerous.

5.3.2 Loss of machine scenario

In order to study which of two machines is critical, we simulated separately, the case of the loss of machine G1 and G2 and followed the transient behavior of the system. As shown in Table 3, in the first case, the two machines are unstable, and the system exhibits a total voltage collapse. While in the second case, only machine G2 becomes unstable, and we note the oscillations of the rotor angle and acceptable electrical power. There are thus among the two machines disturbed during the fault, a machine that is more stable than the other after the elimination of fault. This is confirmed by the necessary critical clearing time that leads to instability. The critical clearing time must be equal to 0.114s for that the machine G2 becomes unstable, while the machine G1 maintains its stability during the definitive loss of G2. The machine G1 does undergo smaller oscillations than the machine G2 in the post-fault configuration. Due to its high nominal power, and that it is chosen as the reference generator. We confirm that machine G1 is critical.

5.3.3 Impact of location of the short-circuit

We will now proceed to study the impact of location of the fault on security margin of the network. A bolted three-phase fault was applied respectively to the nodes 2, 6, 5 and 14 at time t = 250s. The time duration of this fault is 100ms. The choice of nodes is relevant, bus 2 is producer node, a compensator is connected to bus 6, and bus 5 is a consumer node that is linked to two producer nodes and bus 14 is a weak node. Table 4 summarizes all the results of simulations. We noticed that during a fault at bus 2, the system

puts only 0.413s to pass in an extreme state, bus 6 take 5.3s, the bus 5 puts 1.05s and bus 14 passes to an extreme state after 1s.

Let us note that transient system behavior is associated with large oscillations of both machines outputs. Therefore, a short circuit applied at bus 2 and bus 14 impacts greatly the stability of the network. These fault types are therefore classified as a dangerous contingency.

Table 3. State of both machine G1 and G2

Parameters	Loss of G1	State of G2	Loss of G2	State of G1
f(Hz)	40	unstable	49.84	Stable
δ1 (deg)	diverge		15.94	
δ2 (deg)	50		diverge	

Table 4. Variation of Fault Location

Bus	Alert state	Emergency State	Extreme State
2	] 250.1 250.4]	] 250.4 250.413]	] 250.413 250.414]
6	] 250.1 253.7]	] 253.7 255. 3]	] 255.3 260]
5	] 250.1 250.8]	] 250.8 251.05]	] 251.05 251.1]
14	] 250.1 250.8]	] 250.8 251]	] 251 251.3]

5.3.4 *Load increase*

Similarly, we have performed simulations of load variation from 5% to 100%. Table 5 illustrates the stability intervals for three states of the system. We notice that the system satisfies both constraints energy supply and operating for 40% load increase on the system. On the other hand, from 70% the system loses the balance between production and consumption. For a 100% increase in load, the system is subject to voltage collapse. This is explained by the fact that the machine G1, during normal operation, generates only 233MVA (47%) while its nominal capacity is 615MVA. Thus, we note that the machine G1 has reserves to provide the required active power, but obviously, can only satisfy a large load variation.

It is noteworthy that in a multi-machine network, stability for the frequency range is ± 0.4Hz, the critical load of the system is thus 40%. Then, we can indicate that the system is highly vulnerable to voltage collapse in the event of heavy load increase. Furthermore, if we bear in mind that when transition time is small in an emergency or extreme state, this contingency is more severe and requires the treatment.

Numbered contingencies III, V, and VII pose no risk because the network remains stable. As against the others should be analyzed as they can cause loss of synchronism of system. Potentially dangerous contingencies are maintained in a waiting list because some of them can be harmful. Whereas the dangerous contingencies, I, II and IV must be treated. Figure 4 shows the impact of an increase of load on frequency value and rotor angle of the two machines.

Table 5. Stability Margin Depending on Load Increasing

State	Alert State	Emergency state	Extreme State
Stability margin	] 5 40]	] 40 70]	] 70 100]

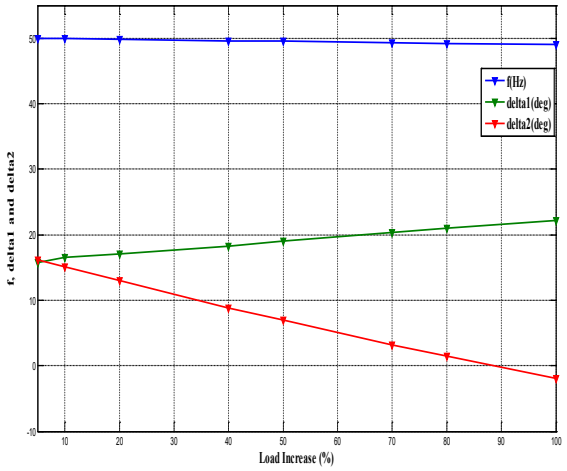


Figure 4. Impact of an increase of load on frequency value and rotor angle of the two machines.

5.4 *Contribution of UPFC on the voltage stability*

The enhancement of voltage stability of the test system using the UPFC is studied. We connect UPFC to the system at different locations. Power dimension of UPFC is ± 200 MVAR. Two scenarios are performed. At first, we consider a total load increase of 40%. Then, a three-phase fault is applied at bus 14. The fault duration is 100ms.

The UPFC was integrated in the middle of line 2-3. Then, it is connected in the middle of line 1-2. Lastly, the UPFC is integrated in the middle of line 9-14.

Table 6 shows the voltage level of all bus under load increase of 40%. Regardless of its location, the UPFC improve almost the voltage level of all bus.

Figure 5 shows the effect of the UPFC on the behavior of voltage at bus 9 and bus 2 with and without the UPFC. The addition of the UPFC has remarkably improved the behavior of voltage.

Here, we consider a bolted three-phase fault at bus 14. The fault duration is 100ms.

Figure 6 and Figure 7 respectively illustrate the temporal evolution of voltages at bus 2 and bus 9 with and without the UPFC.

It should be noted that a short circuit at bus 14 causes a voltage drop which reaches its minimum beyond a certain threshold value and regains its initial value after highly damped oscillations in the presence of the UPFC. Therefore, the UPFC provides the best control of voltage and a good damping of oscillations.

Table 6. Load Increase of 40%

	Sans FACTS	UPFC in line 2-3	UPFC in line 1-2	UPFC in line 9-14
BUS1	1.060	1.060	1.060	1.060
BUS2	1.040	1.040	1.040	1.044
BUS3	1.003	0.985	1.003	1.004
BUS4	1.014	0.992	1.013	1.017
BUS5	1.022	1.003	1.022	1.026
BUS6	0.996	0.982	0.996	1.014
BUS7	1.011	0.992	1.011	1.011
BUS8	1.082	1.067	1.082	1.018
BUS9	0.978	0.959	0.978	1.085
BUS10	0.972	0.954	0.972	0.972
BUS11	0.980	0.963	0.980	0.980
BUS12	0.977	0.962	0.977	0.977
BUS13	0.971	0.956	0.971	0.971
BUS14	0.952	0.935	0.952	0.952

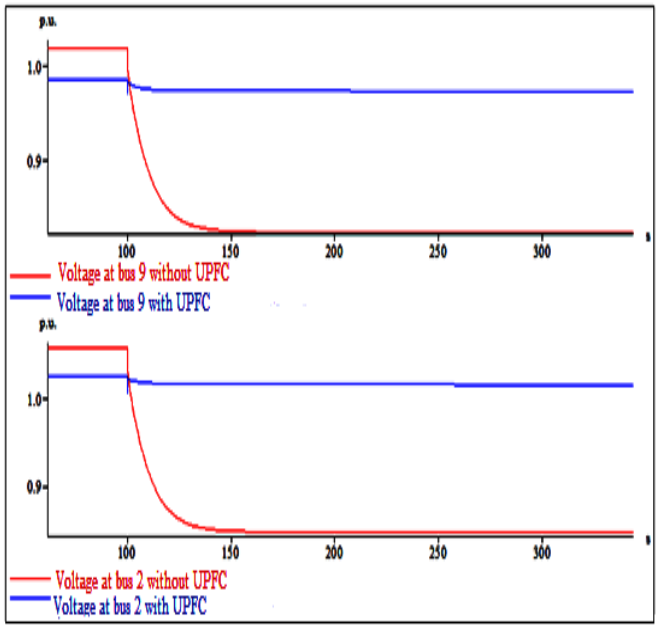


Figure 5 Voltage magnitudes at bus 9 and bus 2 with and without the UPFC.

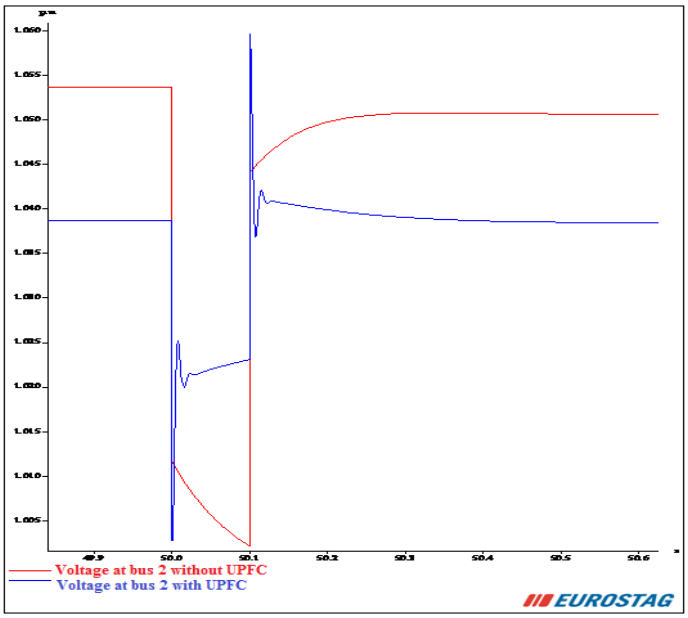


Figure 6 Voltage magnitudes at bus 2 with and without the UPFC.

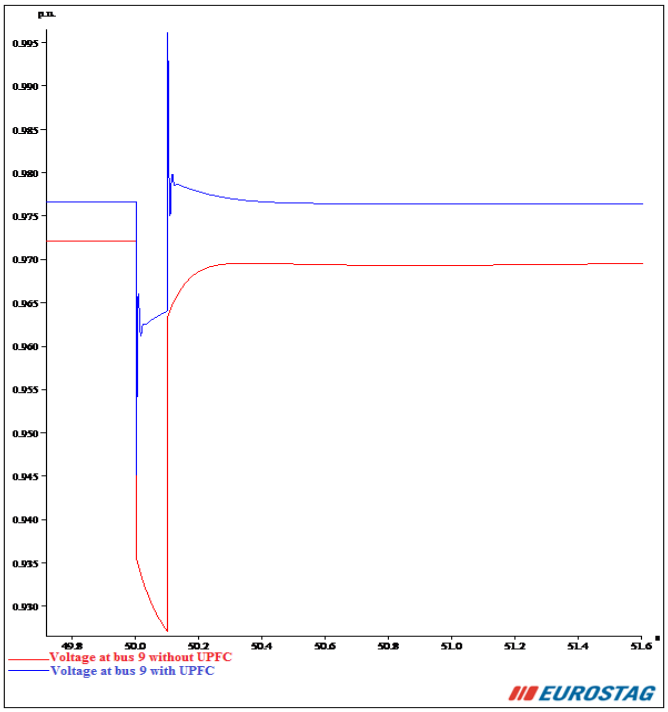


Figure 7 Voltage magnitudes at bus 9 with and without the UPFC.

6. Conclusions

This paper presented a classification of contingencies as well as curative action against faults affecting the voltage stability of the system. Extensive studies of the behavior of IEEE 14-bus system for different contingency cases revealed that is subject to instabilities under well-defined fault conditions. We have demonstrated through this paper that power system suffered from severe contingencies leading to their instability. Thus, we investigated effective methods to improve the stability of power network. A comprehensive analysis of the network security testing has allowed us to classify contingencies in an increasing order of

severity. Some of these contingencies may cause serious instability on the network, and others may not be critical. As a result, it is necessary to focus attention on a limited number of contingencies and seek remedial action through UPFC device, which has a role to reduce the possibility of voltage collapse and limit damage of fault, as overloads and short -circuit. Therefore, the UPFC was most effective in improving voltage stability and reducing the harmful effects of dangerous contingencies.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] K. Alcheikh-Hamoud, "Modeling of large interconnected power systems: application to the safety analysis in a competitive environment," Phd. Engineering Sciences. Institut National Polytechnique de Grenoble - INPG, 2010.

[2] S.r. Gongada ,T.S. Rao, P.Mallikarjuna rao, S. Salima, "Power System Contingency Ranking using Fast Decoupled Load Flow Method," International Conference on Electrical, Electronics, and Optimization Techniques, 2016.

[3] E.Ciapessoni, D.Cirio, S.Massucco, A.Morini, A.Pitto and F.Silvestro, "Risk-Based Dynamic Security Assessment for Power System Operation and Operational planning,", MDPI journal , Energies, 2017.

[4] S. Burada, D. Joshi, and K.D. Mistry, "Contingency Analysis of Power System by using Voltage and Active Power Performance Index," 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, 2016.

[5] Abido, M. A., "Power system stability enhancement using FACTS controllers: A review," The Arabian journal for science and engineering, 34(1B),153-172,2009.

[6] A. Meddeb, H. Jmii, S. Chebbi, "UPFC and SVC Devices for Transient Stability Enhancement,"4th International Conference on Automation, Control Engineering and Computer Science (ACECS) Proceedings of Engineering and Technology – PET, pp.82-87, Vol.19, 2017.

[7] M. Benaissa, S. Hadjeri, S.A. Zidi, "Impact of PSS and SVC on the Power System Transient Stability," Advances in Science, Technology and Engineering Systems Journal, pp. 562-568, Vol. 2, No. 3, 2017.

[8] K.Y. Lee, M. Farsangi, and H.Nezamabadi-pour, "Hybrid of Analytical and Heuristic Techniques for FACTS Devices in Transmission Systems," Power Engineering Society General Meeting, IEEE, 2007.

[9] R.Dubey, S.Dixit, and G.Agnihotri, "Optimal Placement of Shunt FACTS Devices Using Heuristic Optimization Techniques: An Overview," International Conference on Communication Systems and Network Technologies, 2014.

[10] N. Hashim N. Hamzah, M.F. Abdul Latip and A.A. Sallehhudin, "Transient stability analysis of the IEEE 14-Bus test system using dynamic computation for power systems (DCPS)," Third International Conference on Intelligent Systems Modelling and Simulation 2012.

[11] P.K. Iyambo, and R. Tzoneva, "Transient Stability Analysis of the IEEE 14-Bus Electric Power System," Africon conference, IEEE, 2007.

[12] A. Zerigui, "Optimal distribution of electricity production with transient stability constraint established by a statistical approach," Phd. thesis,Univ. Québec, Feb 2015.

[13] X. Tu, L.A. Dessaint, and I. Kamwa, "A global approach to transient stability constrained optimal power flow using a machine detailed model," Can. J. Elect. Comput. Eng., Vol. 36, No. 1, Winter 2013.

[14] R.Kamdar, M.Kumar and G.Agnihortri, "Transient stability analysis and enhancement of IEEE-9bus system," Electrical & Computer Engineering: An International Journal (ECIJ) Volume 3, Number 2, June 2014.

[15] J.C .Chow, R. Fischl, M. Kam, H.H. Yan and S. Ricciardi, "An Improved Hopfield Model for Power System Contingency Classification," Department of Electrical and Computer Engineering Drexel University, Philadelphia, PA 19104, USA.

[16] C. I. Faustino Agreira, C.M. Machado Ferreira, J. A. Dias Pinto, and F. P. Maciel Barbosa, "Application of the Rough Set Theory to the Steady – State Contingency Classification," Power Tech, June 2005, IEEE Russia.

[17] Abdulwahhab A. A. 'Contingency ranking of power systems using a performance index', International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 02 Issue: 02, 2015.

[18] Y. Chen, F.Luo, Y. Xu and Jing Qiu4, "Self-adaptive differential approach for transient stability constrained optimal power flow," IET Gener. Transm. Distrib, pp. 1–10, 2016.

[19] De. Gan, R. J. Thomas, and R. D. Zimmerman, "Stability-Constrained Optimal Power Flow," IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 15, NO. 2, MAY 2000.

[20] F. Mohseni-Kolagar and H. Miar-Naimi , " Transient Analysis of Bang-Bang Phase locked Loops without Cycle Slipping for Frequency Step Inputs," 2012, 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT).

[21] A.Meddeb, H.Jmii, and S.Chebbi, "Heuristic analysis and contingencies classification of case Study IEEE 14-bus," 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp.128–133, 2016.

[22] L. H. Fink and K. Carslen, "Operating Under Stress and Strain," IEEE Spectrum Magazine , 1978

[23] P. M. Anderson and A. A. Fouad, "Power system control and stability," Book, IEEE Press, 1994.

[24] P.Kundur, "Power system stability and control, book" McGraw-Hill,Inc 1993.

[25] M.Pavella, D.Ernest, and D. Ruizvega, "Transient Stability of Power Systems A Unified Approach to Assessment and Control," Book, Kluwer Academic Publishers Boston/Dordrecht/London, 2000th Edition.

[26] Y. Zeng, and Y.Yu,"A Practical Direct Method for Determining Dynamic Security Regions of Electrical Power Systems,", International conf on power system Tech proceeding, IEEE 2002.

[27] P. A. Kaplunovich, K. S. Turitsyn, "Statistical properties and classification of N – 2 contingencies in large scale power grids," 2014 47th Hawaii International Conference on System Science, 2014.

[28] K. Cheong Sou, H. Sandberg, and K. Henrik Johansson, "Electric Power Network Security Analysis via Minimum Cut Relaxation," 2011 50th IEEE Conference on Decision and Control and European Control Conference, Orlando, FL, USA, December 2011.

[29] A. Singh, A. Uddin Ahmad, "Control Reactive Power Flow with UPFC Connected Using Different Distance Transmission Line," Copyright to IJAREEIE DOI: 10.15662/IJAREEIE, 0409034, 2015.

[30] Visakha, K., Thukaram, D., & Jenkins, L, "Application of UPFC for system security improvement under normal and network contingencies," Electric Power Systems Research, 70(1), 46-55, 2004.

[31] Thukaram, D., Jenkins, L., & Visakha, K. , "Improvement of system security with unified-power-flow controller at suitable locations under network contingencies of interconnected systems," IEE Proceedings-Generation, Transmission and Distribution, 152(5), 682-690, 2005.

[32] Gupta, S., Tripathi, R. K., & Shukla, R. D, "Voltage stability improvement in power systems using facts controllers: State-of-the-art review," Power, Control and Embedded Systems (ICPCES), pp. 1-8, 2010.

[33] Kamarposhti, M. A., Alinezhad, M., Lesani, H., & Talebi, N, "Comparison of SVC, STATCOM, TCSC, and UPFC controllers for static voltage stability evaluated by continuation power flow method," In Electric Power Conference, EPEC, IEEE Canada pp. 1-8, 2008.

[34] S.V.R.Kumar, and S. S.Nagaraju, "Transient Stability Improvement using UPFC and SVC," ARPN Journal of Engineering and Applied Sciences, Vol. 2, No. 3, June, 2007.

[35] B.Bhattacharyya, V.K.Gupta, and S.Kumar, "UPFC with series and shunt FACTS controllers for the economic operation of a power system," Ain Shams Engineering Journal, 2014.

[36] X.Ping Zhang, C. Rehtanz, and B. Pal, "Flexible AC Transmission Systems: Modelling and Control," DOI 10.1007/978-3-642-28241-6, Springer-Verlag Berlin Heidelberg, 2012.

[37] S.K. Kodsi and C. A. Canizares, "IEEE 14 Bus system with facts controllers," IEEE Trans, 2003.

[38] Eurostag, Eurostag Software Release Notes, Tractebel-EDF, Release 5.1, Dec 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 412-417 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

The method of correlation investigation of acoustic signals with priority placement of microphones

Bohdan Trembach*,[1], Roman Kochan[1], Rostyslav Trembach[2]

[1]*L'viv Polytechnic National University, Department of Specialized Computer Systems, 79000, L'viv, Ukraine*

[2]*Ternopil Ivan Puluj National Technical University, Department of Automation of Technological Processes and Productions, 46001, Ternopil, Ukraine*

ARTICLE INFO

Article history:
Received: 13 November, 2017
Accepted: 24 January, 2018
Online: 10 February, 2018

Keywords:
Acoustic signals
Correlations
Special processors
Hamming space

ABSTRACT

Examples of analytical calculations of the system characteristics of the hardware and time complexity of the correlation system based on a certain number of microphones and the corresponding number of interrelations are presented. The structural solutions of the hardware special processor implementation of such class of multichannel devices for recognition and identification of types and the spatial location of sources of acoustic signals are developed. The structural model of the spatial identification of sources of acoustic signals in Cartesian coordinates of a two-dimensional Hamming space with the priority placement of microphones as receivers of acoustic signals is proposed.

1. Introduction

This paper is an extension of work originally reported in 14th International Conference the Experience of Designing and Application of CAD Systems in Microelectronics (CADSM) [1].

Development of theoretical foundations of information technology and software - hardware correlation signal processing is actual scientific - applied problem is to be solved in many industries. Identification of sources of acoustic signals (SAS) relative spatial placement of microphones - receivers of acoustic signals (RAS) is also included. The given problem is a primary-industry-technical task of the special technique [2–4].

Analysis of the known research results. In [5–8] the authors created successful but far from optimal solution of such problem is working out localization accumulated information systems. The determination of spatial parameters (θ) azimuth and distance to the SAS (φ) the use of a certain number of (q) correlates for a given number of those (m) microphones is taken as the base of the system.

The example of the structure of Acoustic Localization by Accumulated Correlation (ALAC) system [5] is shown in Figure 1.

It should be noted that the number of required correlators for a given number of chaotic space placed microphones RAS is determined by the condition of symmetry correlation matrix (1).

*Corresponding Author: Bohdan Trembach, L'viv Polytechnic National University, Department of Specialized Computer Systems, 28 a S. Bandera str., 79000, Lviv, Ukraine. Email: trembach.bogdan@gmail.com

$$\begin{array}{ccccc} MK & 1 & 2 & 3 & 4 \\ 1 & 1 & R_{12} & R_{13} & R_{14} \\ 2 & R_{21} & 1 & R_{23} & R_{24} \\ 3 & R_{31} & R_{32} & 1 & R_{34} \\ 4 & R_{41} & R_{42} & R_{43} & 1 \end{array}, \quad (1)$$

where $R_{ii} = 1$, are placed along the diagonal;

$R_{ij} = R_{ji}$ – are identical, according to the symmetry of the matrix (1).

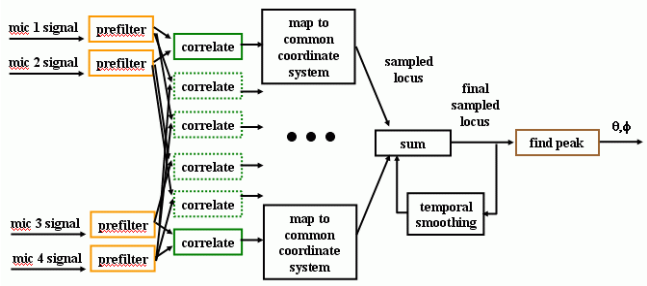


Figure 1. Structure of the ALAC

So, according to the information technology of the correlation processing of the SAS signals in ALAC system, the required amount of (q) correlates for a given number of the (m) microphones is determined by the expression:

$$q = (m^2 - m)/2. \quad (2)$$

The graph of dependency (2) with different numbers of microphones is shown in Figure 2.

For implementation of each correlator in ALAC system the integrated assessment multiplicative correlation function is used by the expression:

$$L(q) = G(\int_{\tau(i,q)-\frac{w}{2}}^{\tau(i,q)+\frac{w}{2}} x_i(t) \times x_j(t-\tau_{i,q}+\tau_{j,q})dt) + \quad (3)$$

$$+ \alpha V_E,$$

where: q – the identifier of the SAS; G – the integrated cross-correlation function; $x_i(t)$ and $x_j(t-\tau_{i,q})$ – current and delayed on the time interval $\pm\tau_{i,q}$ acoustic signals (AS) accordingly; αV_E – damping energy coefficient of the cross-correlation function on the interval $\tau_{i,q}$.

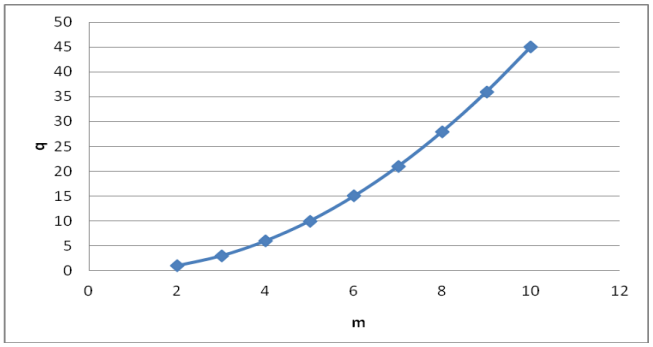


Figure 2. The dependence of the required number of correlators on the number of microphones in the system

The analysis of the analytic expression (3) allows to determine that the implementation of R_{xx} correlates in ALAC system is based on a time delay, $x_i(t)$ multiplying and $x_j(t-\tau)$ integrating analog signals, essentially limiting its functionality, simplifying and increasing the speed and accuracy also prevents it implementation based on digital microelectronics crystals and programmable integrated-circuit logic (FPGA).

The purpose of the work is to develop and explore systemic and structural characteristics of digital special processor computing centered multiplicative correlation function.

2. Formulation of the problem

The purpose of the work is to develop a digital correlator to determine the location of the acoustic signal source.

In order to study the principles of improving and optimizing system features of digital correlators as the basic components of the discovering system, analysis of system features of special processor and computing digital multiplicative correlation estimates by the expression is carried out [9]:

$$R_{xx}(j) = \frac{1}{n}\sum_{i=}^{n} \overset{\circ}{x} \times \overset{\circ}{x}_{i-j}; \quad j \in \overline{0,m}, \quad (4)$$

where: $R_{xx}(j)$ – is centered autocorrelation function; $\overset{\circ}{x}_i$ and $\overset{\circ}{x}_{i-j}$ – are centered digital value and analog signals $x_i(t)$ and $x_j(t-\tau)$; n – is the volume of the sample data set $\left\{\overset{\circ}{x}_i\right\}$; m – is the number of points of correlation function $R_{xx}(j)$; j – is discrete digital delay unit point x_{i-j} in time.

The example of the interaction in time of the digital signals $\overset{\circ}{x}_i$ and $\overset{\circ}{x}_{i-j}$, where C – constant threshold and corresponding asymptotic of the correlation function $R_{xx}(j)$ is shown in Figure 3.

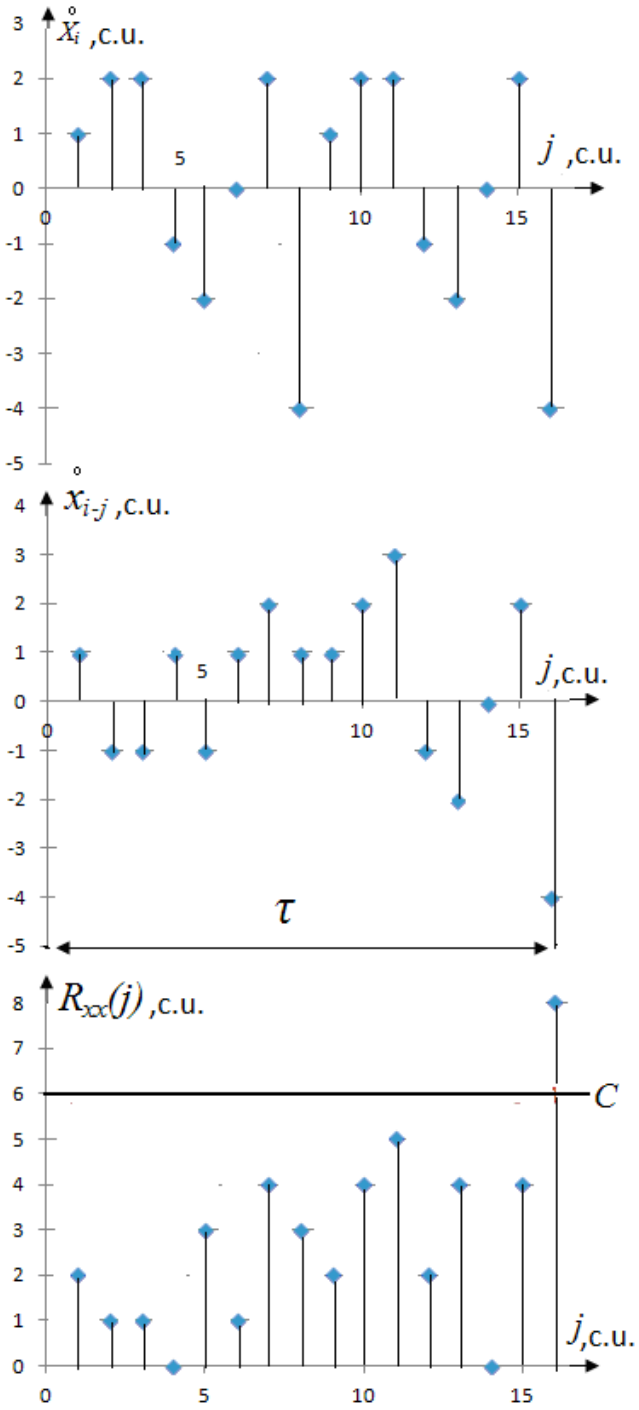


Figure 3. Example of temporal interaction sampled in time and amplitude quantized digital signals and corresponding digital asymptotic estimates multiplicative correlation function in conventional units (c.u.)

Figure 3 shows that at the time of coincidence the current τ digital streams until the moment $\overset{\circ}{x}_i$ of the reception of acoustic signal $x_i(t)$ by a remote microphone SAS level of correlation between the previously obtained a SAS close to the microphone signal $x_j(t-\tau)$ to digital value j – s correlator outputs do not exceed the threshold level constant C and at the time of singing falling at the time fixed main lobe function $R_{xx}(j)$ and accordingly j a numeric value that corresponds to the duration Δt used for direction finding SAS. Obviously, depending on the structure of the stream of digital samples $\overset{\circ}{x}_i$ that reflect the analog signal generated $x(t)$ by a remote SAS in the vicinity of the main lobe $R_{xx}(j)$ functions will emerge sufficiently large lateral lobes, which require separate examination for specific researched objects that have SAS.

In [10] the authors showed that in general for calculation the correlation function $R_{xx}(j)$ according to expression (4) based on a processing arrays of digital data x_i which representing the converted analog-to-digital converter (ADC) input analog signals $x(t)$ you should do the following:

1) define the digital evaluation expectation:

$$M_x = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad (5)$$

2) calculate the array centered values:

$$\overset{\circ}{x}_i = x_i - M_x;\ i \in 1,n; \quad (6)$$

3) calculate the variance:

$$D_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - M_x)^2 = \frac{1}{n}\sum_{i=1}^{n}\overset{\circ}{x}_i^2; \quad (7)$$

4) calculate the correlation function $R_{xx}(j)$ centered on the expression (4);

5) calculate the normalized correlation function:

$$\rho_{xx}(j) = \frac{R_{xx}(j)}{D_x};\ i \in \overline{1,m}; \quad (8)$$

6) perform the comparison of the digital value $\rho_{xx}(j)$ which changes in boundaries $-1 \le \rho_{xx}(j) \le +1$, with boundary constant $0 < C \le 1$:

$$\rho_{xx}(j) > C_0;\ \rho_{xx}(\neq j) < C_0; \quad (9)$$

7) register a digital value j which corresponds to the time delay Δt of the acoustic signal $x(t)$ between two microphones placed at different distance from SAS.

This correlation algorithm of digital processing acoustic signals on a base of multiplicative function (4) is greatly simplified if before the analog-to-digital converter (ADC) pre-differentiation analog $x(t)$ signal is performed and it is passed through a device of automatic gain control.

Such pre-processing of the analog signals $x_i(t)$ and $x_j(t-\tau)$, which are formed on the outputs of the microphones allows to remove from algorithm of calculating operations (5–8) and immediately calculate correlation digital integrated assessment (4) and compare it values at all points j of the constant threshold $R_{xx}(j) > C$.

As it will be shown later, the numerical value of this constant is selected due to the given parameters of digital correlator, which implements the calculation of the multiplicative function $R_{xx}(j)$. Such parameters are:

1) the number of quantization levels of analog signals

$$A = 2^k;\ -\frac{A}{2} \le \overset{\circ}{x}_i \le +\frac{A}{2};$$

2) binary output bit ADC k

$$k = 4,8,10,12;$$

3) the sample size of the database $\left\{\overset{\circ}{x}_i\right\}$

$$n = 2^r;\ r = 4,5,6,7,8,...;$$

4) the number of points correlation function

$$m = 2^l;\ l = 8,10,12....$$

The basic structure of the investigated correlation processing of acoustic signals special processor designed for their processing is shown in Figure 4.

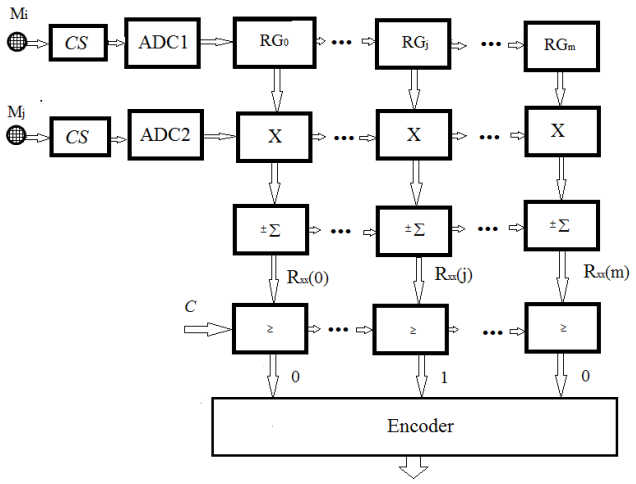


Figure 4. The basic structure of the digital special processor for correlation processing of acoustic signals

The following notions are used in Figure 4: M_i; M_j – RAS microphones; *CS* – conditioners signals; *ADC* – analog-to-digital converter; $RG_0 ... RG_m$ – multi bit shift register source $\pm\overset{\circ}{x}_i$; X – digital binary multiplier; $\pm\sum$ – reversing saving up adder; $\ge$ – module comparing numerical values of the correlation function $R_{xx}(j)$ threshold constant C; encoder – encoder codes Haar – Rademacher.

The work of the digital correlator is as follows. Acoustic signals which are taken by microphone M_i and with a certain time M_j lag are converted to electrical signals which passing

through conditioners signals (CS) are filtered, normalized on power and enter the inputs of the first ADC1 and second (ADC2) alternating analog - digital converters parallel type.

Output k – bit binaries ADC1 come to k – bit input of the multibit shift register (MSR) where RG_j are stored in memory registers, the outputs are fed to the first $j-x$ digital multipliers inputs (X). Binary k – bit codes that are generated at the output ADC2 with a certain time delay corresponding to the time delay of acoustic signals M_j, which are simultaneously fed to all second input ($j-x$) k – bit binary multipliers (X), the outputs of which received $2k$ – bit binaries enter the relevant inputs $(2k+\log_2 n)$ – bit reversible accumulating adder ($\pm\sum$), where codes are received digital values point correlation function $R_{xx}(j)$.

Simplification of operations of division by sample size data set (n) in formula (4) is achieved throughout its multiplicity of 2 degrees, allowing to get the average value of digital codes point correlation function by discarding four junior level in the original binary code accumulating $\pm\sum$. Obtained codes $R_{xx}(j)$ with the bit $2k$, or with less precision $2k–r$, where r = 2,4,6, ... are compared to the corresponding ($j-x$) module compare ($\geq$), the output of which is formed m – bit position code Haar type (00 ... 10 ... 0) , the position "1" of which corresponds to the numerical values of the time delay of acoustic signals received Δt by spatially located microphones M_i and M_j.

For example, when the number of points of the correlation function m = 4096 binary code Δt has 12 bits, that corresponds to Δt = 0.0025 measurement accuracy and uncertainty, and at m = 1024, corresponds to 10-bit binary code and the accuracy does not exceed Δt = 0.001.

The structural scheme shown in Figure 4 can be used for measuring the distance to the SAS by gradient method in case when the source and signal receivers are located anyone line. The example of gradient method is demonstrated in Figure 5.

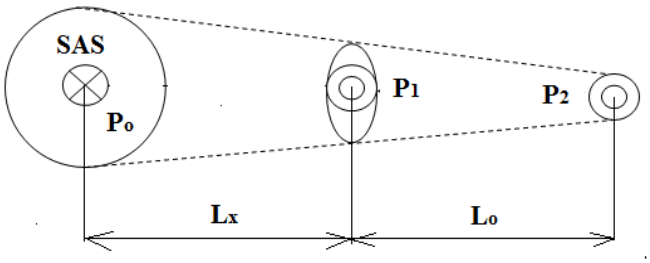


Figure 5. Spatial diagram of SAS capacity changes

The distance to the source is defined by formula

$$L_x = L_0 + \frac{\ln \Delta P}{\alpha},$$

where: α – is the coefficient of acoustic sound energy attenuation in the atmosphere; $\Delta P = P_1 - P_2$ - capacity difference of the signals received from the first and second microphones.

3. Structure of the proposed correlator

The analysis of the expression (4) and the structure of the corresponding digital special processor shows that the presence of the centering operation and the multiplication of alternating digital codes $\pm\overset{\circ}{x}_i \times \pm\overset{\circ}{x}_{i-j}$ complicates the implementation of such a special processor and significantly reduces its performance compared to its functionally equivalent implementation based on the Hamming distance estimation according to the expression

$$G_{xx}(j) = \frac{1}{n}\sum_{i=1}^{n} | x_i - x_{i+j} |, \qquad (10)$$

where centering and multiplication operations are not applied. The application of the Hamming distance estimation based on the calculation of the modular correlation function allows to realize the operation of determining the modular difference between the two digital values $|x_i - x_{i-j}|$ in the microelectronic performance proposed in [9] according.

The purpose of improving the ALAC correlation system is:

1) implementation of the digital representation and correlation acoustic signals processing;
2) reduction of the number of digital correlations of the system;
3) reduction of algorithmic and hardware complexity and increase of speed of digital correlators;
4) implementation of the tabular method for the identification of the spatial location of sources of acoustic signals based on time delay of acoustic signals between microphones;
5) adaptation of the digital correlator to the form of an acoustic signal generated by different sources of sound.

To meet these objective, a method of correlation processing of the acoustic signal with priority placement of microphones [11,12] is proposed on the basis of the multiplicative function (4) converted to the normalized form (8).

Despite successful problem solution for determining time delay of acoustic signals between microphones Δt, such structure of a special processor does not allow to determine independently the spatial coordinates of the source of acoustic signal since it requires more than one base. That is, the number of hardware equipment of such a system increases almost twice. In addition, the use of multiplicative correlation functions $R_{xx}(j)$ and $\rho_{xx}(j)$ requires the implementation of modules for the multiplication and accumulation of reversible interchangeable digital data which considerably complicates the algorithm for processing digital data and the hardware implementation of such a system.

In order to optimize the characteristics of the investigated correlation system the authors propose to implement it on the basis of three microphones, with the priority of spatial placement of one of the microphones at the testbed, and the application of a modular correlation function Gij which allows to identify time delays between acoustic signals Δt_1, Δt_2 and Δt_3 based on the estimation of the Euclidean distance in the Hamming space.

Implementation of the principle of digital processing of acoustic signals allows to reduce significantly the hardware complexity of a single two-channel correlator. The structure of the system is shown on Figure 6.

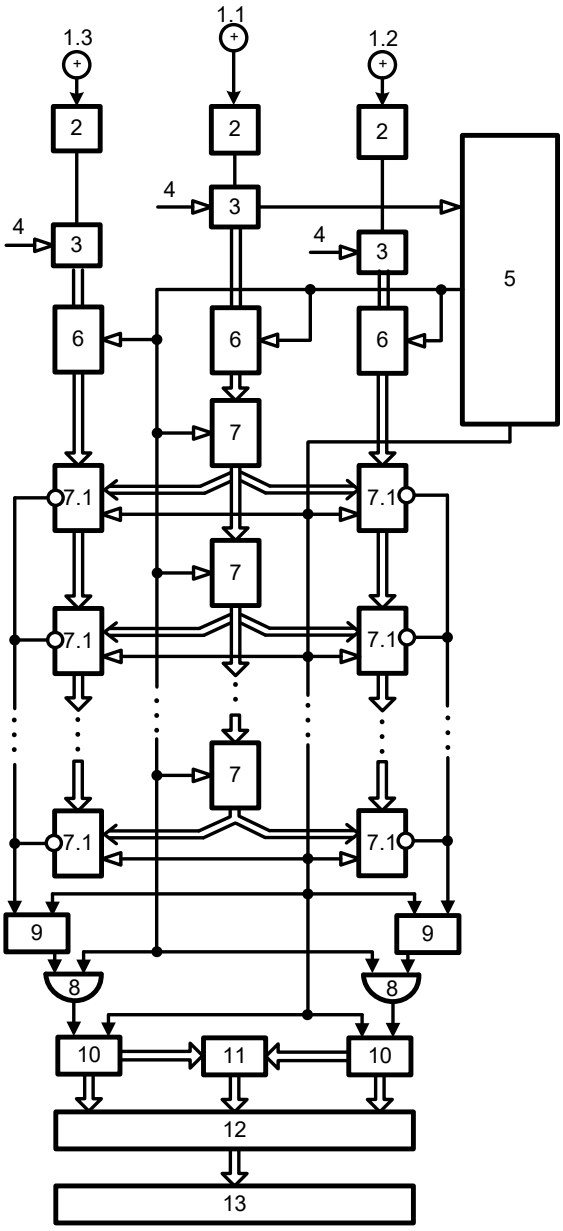


Figure 6. Flowchart of multichannel device for the calculation of modular correlation function

Multi-channel device for calculation of modular correlation function consists of: 1.1, 1.2, 1.3 - respectively: the first priority, second and third receivers of acoustic signals; 2 - amplification automatic adjustment; 3. matched filter of acoustic signals; 4 - reference acoustic signal input; 5 - synchronizer; 6 - parallel-type ADC with output source codes in the binary numeral system of the Rademacher theoretic-numerical basis; 7 - multichannel shift register; ; 7.1 - reverse-accumulation adder; 8 - logical elements AND; 9 - threshold storage of modular differences; 10- RS - triggers; 11 - binary counters; 12 - modular-difference adder; 13 coordinate system based on read-only memory (ROM).

At the beginning of the operating cycle of the device signal S_o of the first output synchronizer 5 generates initial pulse, which resets the memory registers of all cumulative adders of modular differences 9, trigger inputs 10, and binary counters 11 to the zero state.

Input analogue acoustic signals x (t), y (t), z (t), which are generated by a remote source of acoustic signals, enter the input of the acoustic signal receiver 1.1 which is spatially closer to the source of acoustic signals and with some delay in time Δt_1 and Δt_2. Correspondingly, they enter the inputs of related acoustic signal receivers 1.2 and 1.3. Being formed on the outputs of the acoustic signal receivers 1.1, 1.2 and 1.3, electrical signals come to the inputs of the corresponding devices of automatic amplification adjustment 2. Electric signals are produced at their output terminals and are normalized by amplitude and positive indicative potential. Generated, output signals of the automatic amplification adjustment devices 2 flow to the first inputs of matching filters 3. The second filter inputs are connected to the input terminals of reference acoustic signal input 4, and the outputs are connected with the first inputs of corresponding ADC 6.

During the operation cycle of the device the clock signals of the second output of the synchronizer 5 Sx synchronize the formation of source codes x_i, y_i and z_i on the outputs of the corresponding ADCs 6, the corresponding shifts of the digital codes x_{i-j} in the multichannel shift register 7, and pulses coming from the outputs of the corresponding logic elements AND 8 on the inputs of the corresponding counters 10. At the same time, corresponding threshold amounts $\sum_{i=1}^{n}\left|x_{i-j}-y_i\right| \bmod P;\ \sum_{i=1}^{n}\left|x_{i-j}-z_k\right| \bmod P$, s are formed in the accumulated aggregates of the modular differences 9 of the first and second groups, the excess of which causes the formation of a zero potential which converts the corresponding trigger 10 into a single state of the S-input in one of the channels of each group on the inverted outputs of adder accumulator of modular differences 9.

The accumulated amount of impulses in the first counter 11 Δt_1 and in the second counter 11 Δt_2 enters the first and second inputs of the coordinate system 12, and the resulting modular difference in the modular increment adder 12, enters the third input of the coordinate system 13, the output in the form of a code Δt_3. This is the output of the device. Figure 7 shows a timeline diagram of the formation and processing of acoustic signals with the registration of time delays between three microphones Δt_1, Δt_2 and Δt_3 and corresponding values of the source sound coordinates selected from ROM.

For example, at $\Delta L = 1$ м; $i \in \overline{1,256}$, $j \in \overline{1,256}$, which corresponds to the binary codes i and j with 8 bit capacity, the number of sound identified sources with coordinates C_{ij} equals 128^2=8192, and with $i \in \overline{1,64}$ and $j \in \overline{1,128}$ respectively $32 \times 64 = 2048$, and with $i \in \overline{1,16}$ and $j \in \overline{1,32}$ equals respectively $8 \times 16 = 128$. Hence, it corresponds to the spatial dimensions of the testbed targets 256×256; 64×128 and 8×16 m.

4. Research results

The application of multiple autonomous systems of parallel direction finding of acoustic signal sources with respectively

small number of nodes in the Hamming space allows to reduce significantly the accuracy requirements of the digital representation of values Δt_1, Δt_2 and Δt_3.

The required memory for tactical representation of coordinates of the acoustic signal source at $i \in \overline{1,256}$ and $j \in \overline{1,256}$ equals 256 Kbytes.

Generalized chart of acoustic signal processor operation in the Hamming distance based on the modular correlation function is shown on Figure 8.

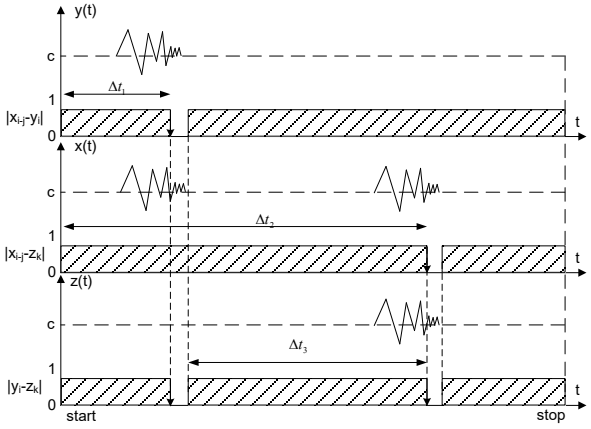


Figure 7. Timeline diagram of code correlation-modular formation of acoustic signal time delay.

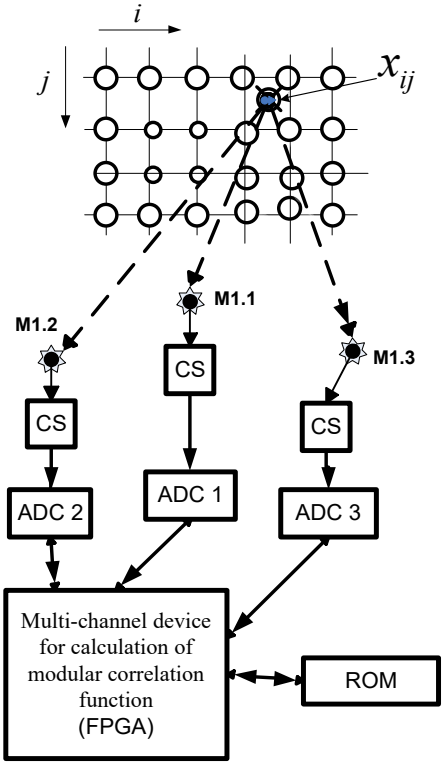


Figure 8. Basic structure of digital special processor for correlation processing of acoustic signals in the Hamming distance

The accuracy of the proposed digital correlator depends on:

- the resolution of a multichannel shift register (Figure 6, element 7);
- the size of square of a Hamming space x_{ij} (Figure 8), which is stored in ROM.

5. Conclusions

The purpose of the work was to develop a structure of a digital correlator to determine the location of the acoustic signal source.

Concept design of an automatic system to determine the location of the acoustic signal source has been proposed. It is based on open architecture and provides connection of multiple autonomous acoustic receivers to the system of correlation special processor using wireless communication channels. This allows automatic collection, processing and data transfer regarding operational conditions in the coverage area of acoustic receivers.

The proposed method of optimizing the structure of multi-channel digital correlates with priority spatial placement of a microphone and application module correlation function to process acoustic signals can significantly simplify the algorithm of calculations, reduce the hardware complexity correlates. This allows enhancing its performance, justifying feasibility and effectiveness of these solutions in the established system of monitoring sources of acoustic signals. The proposed chart allows reducing the number of correlators by 3 times.

Conflict of Interest

No conflict of interest in this paper has been identified.

References

[1] Bohdan Trembach, Roman Kochan, Rostyslav Trembach, "The method of correlation investigation of acoustic signals with priority placement of microphones", in 14th IEEE International Conference on The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Polyana, Ukraine, 2017.

[2] A.M. Krivosheev, V.N. Petrenko, A.I. Prikhodko, Fundamentals of artillery reconnaissance, Sumy State University, 2014.

[3] R. Kochan, B. Trembach, "The concept of distributed automated system of sound artillery intelligence-based provider" Modern information technologies in the sphere of security and defence, 1, 59–63, 2016.

[4] P.E. Trofymenko, Y.G. Filipenko, Sound intelligence station – 100 years, Sumy State University, 3, 198–202, 2009.

[5] S. T. Birchfield and D. K. Gillmor. Acoustic Localization by Accumulated Correlation. [Online]: http://www.ces.clemson.edu/~stb/research/acousticloc/

[6] S. T. Birchfield. A Unifying, "Framework for Acoustic Localization" in 12th European Signal Processing Conference (EUSIPCO). Vienna, Austria, 2004.

[7] S. T. Birchfield and D. K. Gillmor, "Fast Bayesian Acoustic Localization", in the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Orlando, Florida, 2002.

[8] S. T. Birchfield and D. K. Gillmor, "Acoustic Source Direction by Hemisphere Sampling", in the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, USA, 2001.

[9] I. Albanskiy, V. Pikh, T. Zavedyuk, G. Korniychuk, "Theory and Special Processors of Spectral Cosine Fourier Transformation Based on Various Correlation Functions in Hamming Space" in 11th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), Slavske, Ukraine, 2014.

[10] Trembach R.B., Trembach B.R., Sydor A.V., Device of adding multi-bit binary numbers, Ukrainian patent №.117789, 2017.

[11] Bohdan Trembach, Roman Kochan, Rostyslav Trembach, "Multiplex digital correlator with high priority deployment of one of the acoustic signal receivers", Scientific Journal of TNTU, 4(84), 99-104, 2016.

[12] Bohdan Trembach, Roman Kochan, Rostyslav Trembach "Methods of structural design optimization of software hardware problem identification of the spatial parameters of acoustic signals sources", Scientific Journal of KNU, 1(245), 136-139, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 418-425 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Systematic Tool Support of Engineering Education Performance Management

Aneta George*, Liam Peyton, Voicu Groza

School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa ON K1N6N5, Canada

ARTICLE INFO

Article history:
Received: 16 November, 2017
Accepted: 24 January, 2018
Online: 10 February, 2018

Keywords:
Data analysis
Accreditation
Graduate attributes
Performance management

ABSTRACT

Engineering schools must adopt or develop their own systems and processes for graduate attribute assessment. In this paper, we take a systems engineering approach to graduate attribute assessment and propose a system architecture and tool-supported continuous improvement process with key algorithms and mathematical analysis to process the data and provide performance management reporting. Over several iterations, we have introduced and evaluated improved systems support in a collection of tools called the Graduate Attribute Information Analysis system (GAIA). GAIA integrates course indicators as well as external indicators from a variety of sources. It provides a tool-supported continuous improvement process with templates and notifications for all deliverables. There are sound algorithms and tool support and built-in mathematical analysis for data collection and reporting that includes quantitative and qualitative data; weighted grading; historic trend analysis; improved visualization of results; and standardized reports at both the course level and the program level that can be used either for accreditation or to inform program improvement.

1. Introduction

The Canadian Engineering Accreditation Board (CEAB) requires engineering programs to collect data and assess 12 graduate attributes (GA) as part of a continuous improvement process to ensure the programs are adequately preparing students to be licensed as professional engineers in Canada [1]. The CEAB does not specify how graduate attributes should be measured or how the data should be managed. Engineering schools must adopt or develop their own systems and processes. Radloff, de la Harpe, Dalton, Thomas and Lawson report that for over a decade, academic faculty finds GA assessment challenging [2]. They see the need for faculty to develop a shared understanding of how to integrate GA assessment within the teaching of their courses.

This paper is continuation of the work originally presented in 2017 14th International Conference on Engineering of Modern Electric Systems (EMES) [3]. In this paper, we take a systems engineering approach to graduate attribute assessment and propose a system architecture and tool-supported continuous improvement process with key algorithms and mathematical analysis to process the data and provide performance management reporting. Our research methodology to validate our approach is an iterative combination of action and design science research methodology. We work with the faculty of software, computer and electrical engineering programs in the School of Electrical Engineering and Computer Science at the University of Ottawa to perform graduate attribute assessment, while at the same time identifying gaps, and prototyping improved tool support in our lab. Over several iterations, we have introduced and evaluated improved systems support in a collection of tools called the Graduate Attribute Information Analysis system (GAIA).

In our initial assessment of systems support for graduate assessment at uOttawa, we did not encounter the "academic resistance" discussed by Chanock in his article on developing GA assessment criteria [4]. However, we did identify that existing systems, tools and processes already in place at University of Ottawa School of Engineering tended to complicate rather than support the task. The key issues identified were cognitive overload, lack of clarity, and lack of defined systems and processes.

The first version of GAIA was introduced in 2015 and has progressed through three iterations or versions. The architecture integrates course indicators (such as tests, assignments, quizzes, exams or selected questions) from any learning management system (LMS) as well as external indicators from a variety of sources (such as student surveys, employer evaluation or different types of feedback forms). There is a systematic tool-supported

*Aneta George, SEECS, University of Ottawa, Ottawa ON K1N6N5, Canada

continuous improvement process with templates and notifications for all deliverables. There are sound algorithms and tool support and built-in mathematical analysis for data collection and reporting that includes quantitative and qualitative data; weighted grading; historic trend analysis; improved visualization of results; and standardized reports at both the course level and the program level that can be used either for accreditation or to inform program improvement.

2. Literature Review

We researched and compared solutions from two sources - engineering institutions from countries members of Washington accord; and engineering schools across Canada.

There have been several attempts by different universities to create their own tool that will inform student learning, serve accreditation, and inform program development. In 2004, Maxim describes an assessment plan for student performance in three undergraduate engineering programs - computer science, information systems, and software engineering at the University of Michigan at Dearborn,[5]. It lists course learning outcomes (LO), the measurement instruments used to assess them, student achievement on each particular outcome, and the average score [5]. It gained popularity because of its ability to serve simultaneously as a grade book and LO evaluation tool.

The University of West Georgia, US, created a custom-designed software tool to collect, analyze and report assessment data for program requirements and for accreditation purposes. The tool called COMPASS supported an existing open-source classroom management system by adding the ability to map course LO. It allows for review and analysis of collected assessment data, but lacks a direct reporting feature. The data needs to be retrieved and formatted in order to produce a course assessment report. This complicated the process of data analysis and its implementation for informing program improvement [6].

The Curtin University of Technology in Perth, Australia followed a similar approach in developing the Outcomes Database web-based tool in 2005. It mapped course LO, unit LO, generic graduate attributes and assessment rubrics. The reports it generated helped provide an outcomes focused assessment [7]. Although the paper does not specify any method of tool evaluation, it does show that the Outcomes Database was successfully implemented across courses that share common units in Information Technology, Computer Science and Software engineering.

In his thesis, Essa proposes a custom-developed ABET Course Assessment Tool (ACAT) at the University of Nevada Reno that further improves the Michigan approach [8]. The goal of ACAT is to streamline the course assessment process and standardize reporting. To validate the design and user interface of the tool, Computer Science and Engineering faculty members perform a usability study. Usability of the tool was tested, based on International Organization for Standards criteria [9] that measure effectiveness, efficiency and satisfaction in a specified context of use by specific users. The results showed that the tool created is an improvement over the existing manual process used to assess GA. This study is one of the earliest we have identified that attempts to theoretically compare technology adoption issues between three different tools – COMPASS, Outcomes Database and ACAT. A comparison of the features in two off-the-shelf systems, Compass and Outcomes Database, and the ABET Course Assessment Tool (ACAT), shows 100% compatibility for ACAT, 33.3% for Compass and 66.7% for Outcomes Database.

In 2015, the University of Notre Dame, Australia introduced an outcomes-based curriculum mapping system, Prudentia©. It allows for constructive alignment between different learning outcomes and informs assessment and instructional methodology [10]. The week point of the tool as described by the author is its dependence on the quality of the curriculum framework itself.

Deferent statistical ways to analyze and interpret assessment data was presented by the Office of Academic Planning and Assessment (OAPA) at the University of Massachusetts [11].

Integrating data by cohort using analytical methods and correlations, a method similar to the one used in GAIA, is the subject of collaborative research between Northern Arizona University, Christopher Newport University and James Madison University [12].

Weber addresses the major concern in data analysis – reliability of the results [13]. He explores the use of T-test, ANOVA and ANCOVA to compare different treatments of assessment data for reliability.

Approaches by Canadian universities involve adopting suitable vendor products, adapting tools and processes in place and evolving them into their own learning management system that suits the institutional needs.

A comparison of different GA assessment compatible software tools available on the market was provided in a series of papers by Kaupp, Frank and Watts [14, 15, 16]. Considering the ability of each tool to handle and process data, they classify several outcomes-based assessment support vendor products into five categories – Learning Management Systems (LMS), Learning Content Management Systems (L/CMS), Assessment Platforms (AP), Analytics Systems (AS) and Curriculum Mapping Tools (CMT). Their study concludes that none of the tools is able to manage the GA assessment data independently and they can only address a specific aspect of the GA process [17]. Being distinct from the actual assessment process was identified as a common general weakness for all vendor tools. Issues like duplication of grading, need for uploading/re-entering data by instructors, agreement with other tools or systems or data record forms and most of all incompatibility with diverse nature of student assessment turns them into an additional tool rather than major carrier of assessment analysis for program improvement. Furthermore, the authors compiled evaluation criteria for tool adoption identifying that popularity of the tool should not be used as measure for its functionality. The choice should rather be made based on the compatibility of the tool with institutional needs and systems.

Identifying the process of measuring GA performance as the most difficult step in meeting accreditation requirements. Saunders and Mydlarski from McGill University discuss adopting current institutional resources and evolving them into a software information system [18].

Queen's University, University of Calgary, University of Toronto, Concordia University, University of British Columbia, University of Manitoba and Dalhousie University outlined and

compared their institutional approaches to accreditation requirements in a joint publication [19]. In the research, Concordia University is identified as one of the first Canadian engineering schools to develop their own Learning Management System when enabled them to collect data and allow for sharing between users.

Developing our algorithms, we explored the different ways for administering assessment data presented by Carleton University [20], University of Alberta [21] and University of Calgary [22]. Carleton University and University of Alberta merged the twelve CEAB graduate attributes with respective indicators, measures and rubrics. University of Alberta adds an additional step to the process by involving sub-categories associated with learning objectives.

3. GAIA

Engineering programs in Canada need to demonstrate that their graduates possess twelve specific attributes [23]: engineering knowledge; problem analysis; investigation; design; use of engineering tools; individual and teamwork; communication skills; professionalism; impact of engineering on society and the environment; ethics and equity; economics and project management; life-long learning. An Accreditation Board (AB) Report indicates that accreditations completed in 2015 were the first ones, which included Graduate Attribute Assessment and Continual Improvement compliance [24].

In developing the architecture for GAIA we had the following objectives:

- find a way to use (when possible) and/or modify (when needed) assessment tools and rubrics already in place;
- integrate GAs, key performance indicators (KPI), assessment tools, measurement criteria, course information sheets, data collection and analysis into one information system;
- measure GA performance and allow data to inform a continuous improvement process for each program
- generate reports and perform mathematical analysis to inform program improvement;
- be user-friendly and time-efficient;
- minimize cognitive overload for any data collection or analysis task.

3.1. System Architecture

GAIA provides three types of performance management support – collecting data, processing data and generating reports. Figure 1 below shows the GAIA architecture. It includes an academic platform (faculty administrator, program coordinator, program professors, course professors and students) focused on in-class evaluation and a non-academic platform (employers, co-op office, alumni and students) focused on evaluation mechanisms external to the class room. Data collected through the Academic Platform (AP) is associated with program-related courses and traditional courses typically supported by a Learning Management System (LMS). The Non-Academic Platform (NAP) deals with all other sources of data using Registration Management System (RMS), university-run surveys and outside survey sources. GAIA's performance is managed by a system administrator to assure regular assessment data flow and support reporting of results in a timely manner.

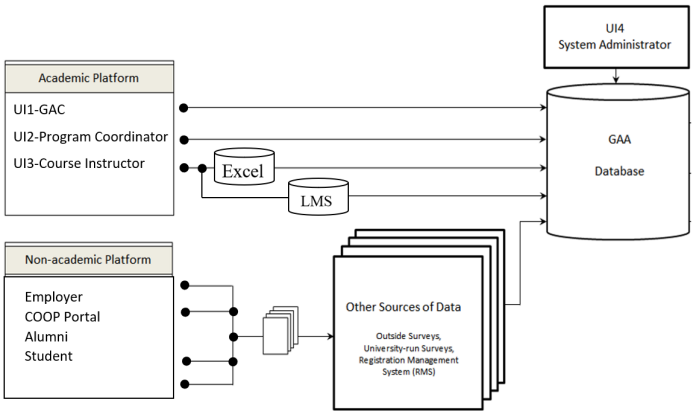


Figure 1: GAIA Architecture

Tool support needs to provide the GAA database (DB) with four types of user interfaces – Graduate Attribute Committee (GAC, UI1), program coordinator (UI2), course instructor (UI3) and system administrator (UI4). Furthermore, it needs to accommodate data fed by different machine interfaces for learning management systems (LMS) and tools already in place for the process of data recording and reporting, as well as qualitative data reported through the COOP portal.

Figure 2 shows the data processing performed by GAIA. It uses different algorithms to process graduate attribute assessment data and generate reports for accreditation agencies in addition to supporting a continuous improvement process for each engineering program. It applies different algorithms to process qualitative and quantitative data input by faculty directly, or imported from a variety of sources. We have identified four MI sources for GA assessment data – LMSs, COOP portal, employer evaluation reports and student surveys. GAIA's architecture pulls external data from these sources in three different ways:

- Using CSV files
- SQL server
- ODBC compliant sources

GAIA's architecture also supports Excel's External Data feature. Once data is initially placed at the location selected by the user, the Refresh button from the Excel Data tab allows for automated updates reflecting data changes at the source. This process is usually set up and performed by the system administrator. GAIA also allows the use of Open Database Connectivity (ODBC) as a programming interface to pull data from different database management systems (DBMS) such as Blackboard or D2L, both currently used at the University of Ottawa.

3.2. Tool Supported Continuous Improvement Process

GAIA supports a systematic continuous improvement process for each program. Creating and implementing improvement steps are initiated and overseen by the members of respective Program Curriculum committees. Success is reported and gaps identified based on consecutive cycles of data analysis performed supported by GAIA as shown in Figure 3 below.

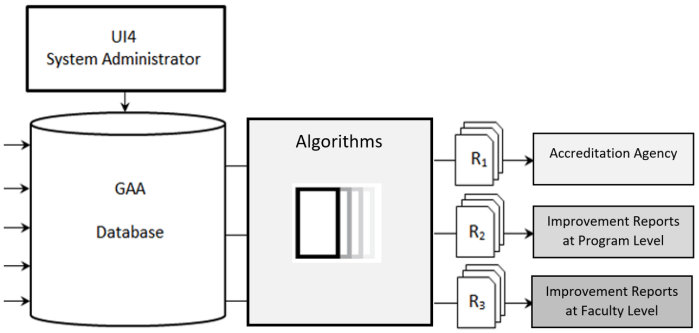


Figure 2: Data Processing for Reporting

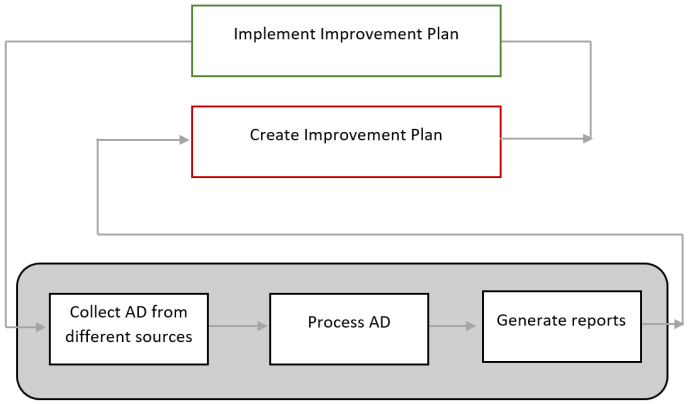


Figure 3: GAIA-supported Continuous Improvement Cycle

The top-down approach, illustrated in Figure 4, mandated by a common continuous improvement process (CIP) leads to a common set of indicators that can be shared across programs. There is a standardized program report that charts a common set of indicators across the 12 graduate attributes. The data for those indicators is obtained from measuring student achievement on the learning outcomes compiled into the common set of high-level indicators specified in the Program Reports. Initially, each program had its own set of performance indicators, which was problematic when trying to achieve consistent reporting at the faculty level that could support cross-program comparisons. Each program aggregated data is clustered into three levels:

- Level I (Course Level): Learning Outcomes presented at the bottom row in Figure 4;

- Level II (Program Level): Performance Indicators, indicated in Figure 4 in black ovals;

- Level III (Meeting Accreditation Requirements): Graduate Attributes analysis for accreditation and program improvement.

At level III we have the 12 graduate attributes specified by CEAB for graduate attribute assessment. Each program has its own set of measurement criteria used to report on achievement for a particular performance indicator (KPI). Selected KPIs are integrated to report on respective graduate attributes (GA). Those indicators are determined by the measurement of learning outcomes for particular courses. The compilation into indicators is mandated in a standardized fashion by the faculty to ensure that achievement is reported in a consistent fashion within a standardized process of continuous improvement.

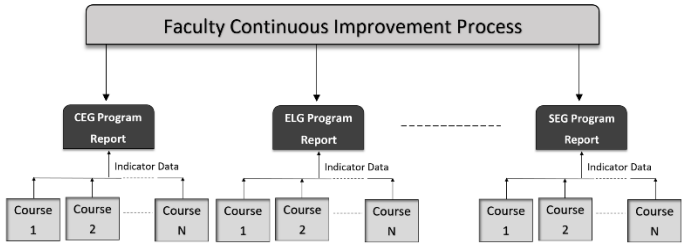


Figure 4. Faculty Continuous Improvement Process.

3.3. Tool-Supported Data Collection and Reports

GAIA provides data collection and reports at two different levels using Course Data Entry Form (CDEF) for individual courses and the Program Report Form (PRF) for an entire engineering program. Data for CDEF is provided by UI3 or input through CVS, SQL server or ODBC compliant sources. Data to Data for PRF is automatically processed and integrated from all the CDEFs for a program. CDEF provides reports and analysis for individual courses. PRF provides reports and analyses across the entire program.

As discussed in section 3.1, GAIA generates three major types of reports – reports for accreditation, improvement reports at program level and improvement reports at faculty level. They are presented in table and graph forms to improve visibility and usefulness. Different tool components, algorithms and data sources are involved in each.

A compiled data report is generated at course level and reflects GA achievement for courses which typically take two semesters to complete. Example for such course is a capstone project. Figure 5 and Figure 6 show a sample report generated for the purpose of this paper in table and graph form respectively.

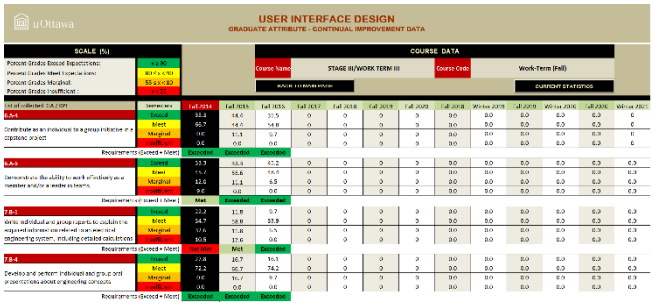


Figure 5. CDEF Table Report Form.

As illustrated in Figure 5 above the report uses color coding to better indicate the level of achievement for the course toward accreditation requirements. Using the CEAB meeting requirements scale with 60% - 80% boundaries. All courses that show achievement per GA below the lower bound did not meet expectations. Their status is shown in red. Above the upper bound of 80% indicates exceeding expectations (shown in dark green shading). The accepted level for meeting accreditation requirements is indicated in light green.

The current statistics graph illustrated in Figure 7, provides a graphical presentation of a historic trend of data collected for selected course. It allows for easy comparison of achievement for

an entire accreditation cycle and gives a better visualization of the overall graduate attribute assessment against meeting accreditation requirements. The Report button allows users to see a historic trend of data in table form.

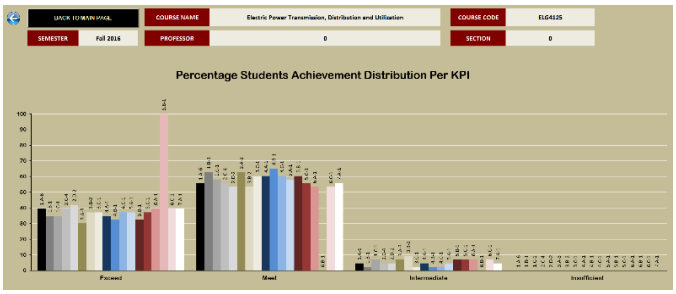


Figure 6. CDEF Graph Report Form.

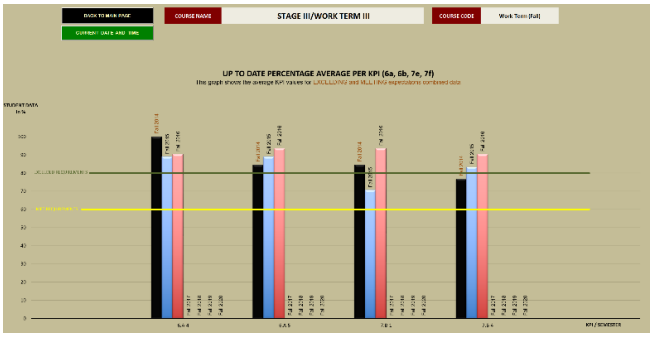


Figure 7. CDEF Current Statistics Report Form.

The Program Report Form (PRF) provides similar statistics which reflect achievement at program level. These reports are used by the curriculum committee members to inform program improvement. The PRF is a read-only workbook. It collates the assessment data provided by CEDF for all courses on a semester basis. This component of the tool measures and reports the cumulative impact individual courses have on overall program performance. It is generated in table and graph form per graduate attribute. A sample report is shown on Figure 8.

(a) (b)

Figure 8. PRF generated report (a) table form, (b) graph form.

A special report form is used to track COOP work-placement data. Data is fed into GAIA from the University of Ottawa COOP Portal using CVS, SQL server or ODBC compliant sources. The report is used to inform on students' ability to secure their first COOP position. A sample of such report form is simulated for the purpose of this paper and is shown on Figure 9.

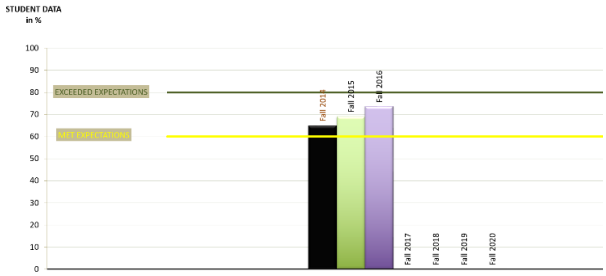


Figure 9. Work-Term Placement Data Report Form.

GAIA allows for comparison between cohorts by generating two reports: the Graduate Attribute Report per Cohort (GAR/C) and the Course Progression Report per Cohort (CPR/C). The former arranges averaged GA data per attribute, while the latter tracks students' achievement as student progress in their program. Data from both cohort reports are used for comparison to provide a historic data trend for further curriculum development. Furthermore, the COOP Progress Report per cohort (COOPR/C) adds reliability in analyzing students' employability and professional skills assessment provided by employers. Table 1, Table 2 and Table 3 below illustrate the datum of cohort reports. It is based on SEG course sequence selected by the program GA Committee to report on GA achievement. Color codes indicate the relative level of the course within the program as follows:

- blue shading is used for Year 5 courses
- green shading is used for Year 4 courses
- peach shading is used for Year 3 courses
- magenta shading is used for Year 2 courses
- no shading is used for Year 1 courses

Table 1. Graduate Attribute Report per Cohort (GAR/C)

	Graduate Attribute											
Courses	GA1	GA2	GA3	GA4	GA5	GA6	GA7	GA8	GA9	GA10	GA11	GA12
	SEG 3101	SEG 3101	SEG 3103	SEG 491 1	SEG 2105	SEG 2105	SEG 2911	SEG 491 1	SEG 2911	SEG 2911	SEG 4105	SEG 4911
		SEG 491 1			SEG 3101	SEG 4105	SEG 491 1	SEG 1911				SEG 1911
					SEG 3102	SEG 491 1						

Table 2. Course Progression Report per Cohort (CPR/C)

	Academic Year				
Course	Year 1	Year 2	Year 3	Year 4	Year 5
	SEG1911	SEG2105	SEG2106	SEG3101	SEG4911
		SEG3103	SEG2911	SEG3102	SEG4105
		SEG3125		SEG4145	

Table 3. COOP Progress Report per Cohort (COOPR/C)

Academic Year				
Year 1	Year 2	Year 3	Year 4	Year 5
SEG1911	COOP Placement I	SEG2901	SEG3901	SEG3902

To illustrate the mechanism of data sorting behind the three cohort reports we generated random data that is being used for demonstration purposes in all tables and graphs. To follow the course progression within a cohort GAIA combines a historic trend data per respective courses according to the course sequence of the specific engineering program (i.e. software engineering in this sample). To generate GAR/C GAIA for cohort N, GAIA will use

Year (N) data for SEG4911 and SEG4105, Year (N-1) data for SEG3101, SEG3102 and SEG4145; Year (N-2) data for SEG2106 and SEG2911; Year (N-3) data for SEG2105, SEG3103, SEG3125; and Year (N-4) data for SEG1911. This is illustrated in Table 4 below.

Table 4. Use of historic trend data for generating cohort reports (GAR/C and CPR/C)

Cohort (Year)	Course	Graduate Attribute
GA Data (N)	SEG4911	2, 4, 6, 7, 8, 12
	SEG4105	6, 11
GA Data (N – 1)	SEG3101	1, 2, 5
	SEG3102	5
	SEG4145	11
GA Data (N – 2)	SEG2106	Xxx
	SEG2911	7, 9, 10
GA Data (N – 3)	SEG2105	5, 6
	SEG3103	3
	SEG3125	Xxx
GA Data (N – 4)	SEG1911	8, 12

The data-supported framework for cohort reports is illustrated in Figure 10. It represents a combination of methodologies described in Figure 1, Figure 2, Figure 4 and Table 4.

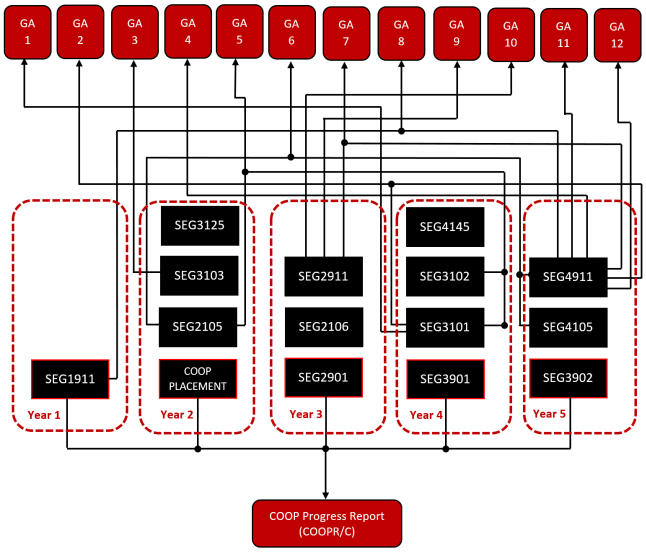


Figure 10. Data-Support Framework for Cohort Reports.

4. Results and discussion

GAIA's improvement and further development is based on ongoing research on latest developments in the area of data management and analytics and the input provided by its users. Mapping the goals stated in section 3 has enabled us to establish an ongoing user-centered tool evaluation. It addresses the different users (U1, U2, U3 and U4 as listed in Figure 1) targeting their specific context of use. Thus, GAIA needs to reflect specific user requirements and specific program requirements for each

university actor. The GAA DB needs to accommodate different types of input data, analyze qualitative and quantitative data, and produce reports has to meet the latest accreditation requirements and inform a continuous program improvement process. Input on the report efficiency and efficiency of the improvement process are being provided by the GA Committee members per program. The ongoing evaluation process as described above is illustrated in Figure 11.

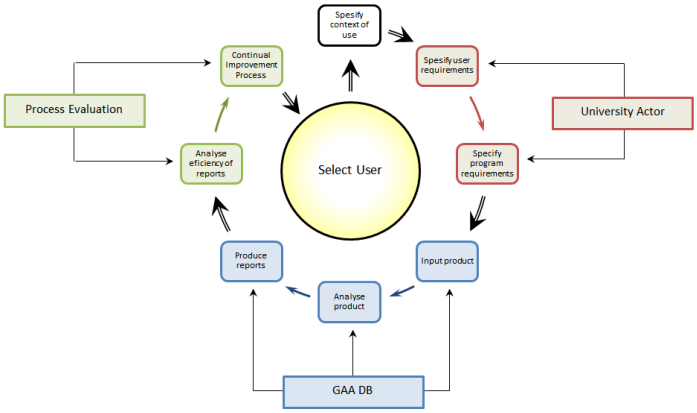


Figure 11. User-Centered Tool Evaluation Process

The following list of evaluation criteria are used for tool evaluation:

4.1. Perceived Ease of Use (PEOU)

This is a core variable, which measures the degree of belief that mastering the tool will not require any extra effort. In the TAM 2 model this construct is defined as a direct determinant of perceived usefulness [25]. Example for its use in academia is the study performed by Park in 2009 [26].

4.2. Perceived Usefulness of the Tool (PU)

Perceived Usefulness is the degree to which a user believes the tool will help successfully complete the task and excel at their job performance. This is a core variable, used to measure adoption in academia assessing learning performance, academic productivity and supporting learning process [26].

4.3. Attitude Toward Using the Tool

The attitude toward use measures the user's feeling about performing the task using the tool. It shows the user's desire to actually use the tool, their positive or negative evaluation of performing the behavior. It measures the ability to perform the task faster, improve user performance when using the tool, using the tool is related to increase of productivity and effectiveness. As a core variable, it is a part of almost every TAM construct set. Samples for its use in academia can be found in the studies by Kim, Park and Tsai [26, 27, 28].

4.4. Behavior Intention to Use the Tool (BI)

It measures the strength of the user's intention to use the tool or the degree of one's willingness to use the tool. It is one of the best indicators of the real usage of the tool. In other words, the actual use depends on the user's intentions to apply effort. It is a combined measure of the wish to finish the task and planning its use in the future.

4.5. Perceived Usage of the Tool

This is the amount of time interacting with the tool and the frequency of its use. Measuring this variable has highest

importance for Faculty Administrator (UI1) as manager interested in evaluating the impact of the tool as a whole.

4.6. Experience Using the Tool

Prior experience was identified as a determinant of behavior in 1980 by Ajzen and Fishbein long before Davis proposed the technology acceptance model [29]. Measuring the variable is mentioned in several studies [30, 31]. According to the studies, experienced users show strong correlation to perceived usefulness of the tool and the behavioral intention to use the tool. Being related to the number of years using the tool, this construct provides valuable information in terms of university users being ready to deal with measuring graduate attributes an ongoing process in a long term.

4.7. Complexity of the Tool

This variable measures the extent to which the user expects to use the tool without any additional effort. It has to do with the tool being difficult to understand and use. Complexity is measured in terms of time taken for the task and integration of tool usage results into existing tool. Complexity is inversely proportional to perceived usefulness (PU) and perceived usage as aspects of adoption. It is also measured by the extent to which the user realizes the possibility of computer crash or data loss.

4.8. Mandatory versus Voluntary Use of Tool

This variable is measured by the extent to which adoption is perceived as a mandatory or non-mandatory task. It is positively related to the behavioral intention to use the tool. In our particular case, the use of the tool is mandatory (requirements enforced by CAEB), so definitely imposed on Faculty Administrator (UI1) and Program Coordinator (UI2). Therefor this variable is not-applicable for them and should be noted as such in their respective evaluation criteria. Participation is mandatory for UI1 and UI2. It becomes mandatory for UI3 if their course is included. At the same time, UI3 cannot easily be coerced for tenured professors so UI3 adoption is critical. Similarly, indirect UI for students, employers etc. cannot be coerced so adoption into process is also critical. The point of this is how to weight the importance of various aspects of the tool. The reporting to CEAB is mandatory and critical so that UI adoption is both constrained and important, but adoption by administrator and coordinator will be coerced so not so critical.

4.9. Interoperability

This measure the tool compatibility with other systems or tools used simultaneously.

4.10. Handling Assessment Data

This measure reflects on tool's ability to serve as database and allow for historic trends of analysis and comparison. The ability to use quantitative and qualitative data is to be evaluated as well.

4.11. Reporting Ability

Tests the tool against its ability to produce reports on course and program levels. Quality of reports and their use to inform program improvement is part of this criteria.

4.12. Alignment with CEAB accreditation requirements

This construct provides direct information on the user's belief that the work performed will serve the need it was intended for. Analysis of the results will have a wide range of application – from indication for improving the tool, to explanation about the user's attitude and perceived intentions

5. Conclusions and Future Work

Creation of GAIA reflected the need for our engineering programs to respond to CEAB accreditation requirements. Ever since it follows the changes implemented to the criteria to provide best support for reporting graduate attributes. Flexibility of the tool and the fact that it is onside made allows for immediate modifications to take place. It is constantly improved following the requests and recommendations from instructors, program coordinators or faculty administrators as well.

In future work, a structural equation model showing the relation between major and external variables in terms of hypotheses will be added. Such a graphical representation of the research model will clearly show the relationships between constructs and specify hypotheses with respect to those relationships. The structural equation modeling (SEM) technique can be employed using the LISREL program. Then direct/indirect effect and t-values will have to be calculated to identify the state of each hypothesis as "Supported" or "Not-supported". Secondary future research will also include further improvement of the model and and more case studies to validate. The goal will be to achieve technology adoption results consistent with expectations. Such results will allow us to measure the tool satisfaction level and predict its usability in specifically constrained contexts.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to acknowledge funding from NSERC and the School of Electrical Engineering and Computer Science at the University of Ottawa which supported this research. We would especially like to thank Claude D'Amours, Emad Gad, Voicu Groza, Stéphane Somé and Marc-André Daoust for their partnership and support.

References

[1] Engineers Canada, "Canadian Engineering Accreditation Board Accreditation Criteria and Procedures," 2014, pp. 1-118. Engineers Canada, Ottawa, ON, K2P2K3, ISSN 1708-8054

[2] A. Radloff, Barbara de la Harpe, Helen Dalton, Jan Thomas, Anne Lawson, "Assessing Graduate Attributes: Engaging Academic Staff and their students," in ATN Assessment Conference 2008: Engaging Students in Assessment, 2008, p. 90.

[3] A. George, L. Peyton, V. Groza, "Graduate Attribute Information Analysis System (GAIA) – From Assessment Analytics to Continuous Program Improvement - Use of Student Assessment Data in Curriculum Development and Program Improvement" in 14th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania 2017. https://doi.org/10.1109/EMES.2017.7980398

[4] K. Chanock, "Developing criteria to assess graduate attributes in students' work for their disciplines," J. Learn. Dev. High. Educ., no. 6, p. 16, 2013.

[5] Bruce R. Maxim, "Closing the loop: assessment and accreditation", Journal of Computing Sciences in Colleges, vol. 20, issue 1, October 2004, pp. 7-18.

[6] Abunawass A., Lloyd W., Rudolph E., "COMPASS: a CS program assessment project", ACM SIGCSE Bulletin, vol. 36, issue 3, September 2004. Pp. 127-131, NY, USA. {DOI: 10.1145/1026487.1008031}.

[7] Konsky B., Loh A., Robey M., Gribble S., Ivins J., Cooper D., "The Benefit of Information Technology in Managing Outcomes Focused Curriculum Development Across Related Degree Programs", Eighth Australasian Computing Education Conference (ACE2006), Hobart, Tasmania, Australia, January 2006.

[8] Eugene O. Essa, "ACAT: ABET Course Assessment Tool," University of Nevada, Reno, 2010, 73 pp., 2010. Available as of Apr. 19, 2017 from http://csi.fau.edu/wp-content/uploads/2013/03/Essa_ACAT_MS-Thesis_U-of-Nevada_Reno_2010.pdf

[9] ISO 13407, "Human-centered design processes for interactive systems", International Standards, Reference number ISO 13407:1999(E), www.isostandards.com.au. Summary online. Accessed July 10, 2016 from http://www.ash-consulting.com/ISO13407.pdf

[10] Steketee C., "Prudentia©: An outcomes curriculum mapping system", Teaching and Learning Forum 2015 - Teaching and Learning Uncapped, Category Professional Practice, pp. 9, 2015.

[11] M. Stassen, K. Doherty and M. Poe, "Program-based review and assessment. Tools and technoques for program improvement", Umass Amherst, pp. 1-62, Fall 2001

[12] S. L. Pieper, K. H. Fulcher, D. L. Sundre and T. D. Erwin, "What do I do with the data now? Analyzing assessment information for accountability and improvement", Research and Practice in Assessment, vol. 2, Issue 1, pp. 1-8, Jan. 2008

[13] E. Weber, "Quantifying student learning: How to analyze assessment data", Ecology 101, pp. 501-511, October 2009

[14] J. Kaupp, B. Frank, and C. Watts, "Evaluation of software tools supporting outcomes-based continuous program improvement processes," in 2013 Canadian Engineering Education Association (CEEA13) Conf., 2013. https://doi.org/10.24908/pceea.v0i0.4861

[15] Jake Kaupp and Brian Frank, "Evaluation of Software Tools Supporting Outcomes-Based Continuous Program Improvement Process. Part II," Canadian Engineering Education Association (CEEA14) Conf. 2015. https://doi.org/10.24908/pceea.v0i0.5856

[16] J. Kaupp and B. Frank, "Evaluation of software tools supporting outcomes-based continuous program improvement processes. Part III," in 2015 Canadian Engineering Education Association (CEEA15) Conf., 2015. https://doi.org/10.24908/pceea.v0i0.5764

[17] K. Chanock, "Developing criteria to assess graduate attributes in students' work for their disciplines," J. Learn. Dev. High. Educ., no. 6, p. 16, 2013.

[18] A. L. Saunders and L. B. Mydlarski, "Evolution of graduate attribute assessment and continuous program improvement in the Faculty of Engineering at McGill University," in 2015 Canadian Engineering Education Association (CEEA15) Conf. 2015. https://doi.org/10.24908/pceea.v0i0.5763

[19] J. Kaupp, B. Frank, R. Brennan, S. Mccahan, L. Narayanan, P. Ostafichuck, N. Sepehri, and K. C. Watts, "A Comparison of Institutional Approaches to CEAB Graduate Attribute Requirements" in Canadian Engineering Education Association (CEEA12) Conf., 2012. https://doi.org/10.24908/pceea.v0i0.4642

[20] J. Harris, D. Russell, and A. Steele, "Progress on Defining the CEAB Graduate Attributes at Carleton University, in Canadian Engineering Education Association Conference, 2011. http://dx.doi.org/10.24908/pceea.v0i0.3628, pp. 1-5, 2011

[21] E. Csorba, D. Chelen, N. Yousefi, N. Andrews, and C. More, "Graduate Attributes at the University of Alberta," a report of the Committee on the Learning Environment (CLE) Subcommittee on Attributes and Competencies, http://www.provost.ualberta.ca/en/~/media/provost/Documents/Information/GraduateAttributes.pdf, pp. 1-22, June 5, 2013

[22] "UC Graduate Attributes Framework", Faculty of Science, U of Calgary, http://www.ucalgary.ca/science/teachinglearning/graduate_attributes_framework, 2017

[23] CEAB, "2014 Canadian Engineering Accreditation Board Accreditation Criteria and Procedures", pp. 118, ISSN 1708-8054, Engineers Canada, 2014. Online "https://engineerscanada.ca/sites/default/files/2014_accreditation_criteria_and_procedures_v06.pdf. Accessed February, 2018.

[24] Iachiver, G., "Accreditation Board Report, EC Board Meeting – September 30, 2015, https://engineerscanada.ca/sites/default/files/AB-Report-September-2015.pdf

[25] Venkatesh V. and Davis D., "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies", Management Science © 2000 INFORMS, vol. 46, no. 2, Feb., 2000, pp.186-204.

[26] Park S., "An Analysis of the Technology Acceptance Model in Understanding University Students' Behavioral Intention to Use e-Learning", Educational Technology & Society, 2009, vol. 12, issue 3, pp. 150–162.

[27] Kim Y., Chun J., Song J., "Investigating the role of attitude in technology acceptance from an attitude strength perspective", International Journal of Information Management, Volume 29, Issue 1, February 2009, Pages 67–77.

[28] Tsai Wen-Chia, "A study of consumer behavioral intention to use e-books: The Technology Acceptance Model perspective", Innovative Marketing, Volume 8, Issue 4, 2012.

[29] Ajzen I., Fishbein M., "Understanding attitudes and predicting social jj..abehavior", Prentice Hall, 1980, pp. 278. {ISBN: 0139364439, 9780139364433}.

[30] Yang, H.-d., and Yoo, Y. "It's all about attitude: revisiting the technology acceptance model," Decision Support Systems, vo.38, issue 1, Oct 2004, pp. 19-31.

[31] Alharbi S., Drew S., "Using the Technology Acceptance Model in Understanding Academics Behavioural Intention to Use Learning Management Systems ", (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 5, No. 1, pp. 13, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 426-433 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Co-designed accelerator for homomorphic encryption applications

Asma Mkhinini[1,2,3*], Paolo Maistri[1], Régis Leveugle[1], Rached Tourki[3]

[1]*Univ. Grenoble Alpes, CNRS, Grenoble INP[**], TIMA, F-38000 Grenoble, France*

[2]*Univ. of Monastir, EμE, 5019 Monastir, Tunisia*

[3]*Univ. of Sousse, Eniso, BP 264 Erriadh 4023, Tunisia*

ARTICLE INFO

Article history:
Received: 17 November, 2017
Accepted: 16 January, 2018
Online: 10 February, 2018

Keywords:
Homomorphic encryption
Hardware accelerator
Modular polynomial multiplication
High Level Synthesis

ABSTRACT

Fully Homomorphic Encryption (FHE) is considered as a key cryptographic tool in building a secure cloud computing environment since it allows computing arbitrary functions directly on encrypted data. However, existing FHE implementations remain impractical due to very high time and resource costs. These costs are essentially due to the computationally intensive modular polynomial multiplication. In this paper, we present a software/hardware co-designed modular polynomial multiplier in order to accelerate homomorphic schemes. The hardware part is implemented through a High-Level Synthesis (HLS) flow. Experimental results show competitive latencies when compared with hand-made designs, while maintaining large advantages on resources. Moreover, we show that our high-level description can be easily configured with different parameters and very large sizes in negligible time, generating new designs for numerous applications.

1. Introduction

This paper is an extension of the work originally presented in 2017 IEEE International Symposium on Design and Diagnostics of Electronic Circuits and Systems [1].

Homomorphic encryption is one of the most significant advances in cryptography in the last decade. It allows arbitrary computations on ciphertexts without compromising the content of the corresponding plaintexts. Thus, data can remain confidential while it is processed, enabling useful tasks to be accomplished with data being stored in untrusted environments. Considering the recent growth in the adoption of the cloud computing and the large deployment of the internet of things, homomorphic cryptography will have a major impact on preserving security and privacy in the coming years. Enterprise customers in the medical and financial sectors, for example, can potentially save money and streamline business processes by outsourcing not only the storage but also the computation of their data to public clouds.

Since the introduction of the first fully homomorphic encryption (FHE) scheme by Gentry [2] in 2009, we have noticed substantial research in the area, for the purpose of designing new homomorphic encryption algorithms, improving the schemes, their implementations, and their applications. Among them, the schemes that are based on Ring Learning With Errors (RLWE) [3] [4] [5] are among the most efficient homomorphic schemes because of their simpler structure, strong hardness assumptions, reduced key size, and reduced ciphertexts expansion with respect to previous schemes. Many of these RLWE-based schemes have been implemented in software [6] [7] [8] [9]. Results report very large latencies and resources consumption. So, in order to improve the performance of homomorphic encryption schemes, there has been research into the hardware acceleration of various homomorphic schemes and their building blocks. To date, there have been few hardware implementations for cryptosystems based on the RLWE problem. The corresponding architectures were mainly designed for fixed and small length operands and optimized for a restricted set of parameters [10], [11], [12], making them limited in terms of target applications and security requirements.

This paper presents a flexible and configurable accelerator implementing modular polynomial multiplication; the main performance bottleneck in RLWE-based homomorphic schemes. The work describes a software/hardware (SW/HW) co-designed architecture based on a High-Level Synthesis (HLS) approach. By combining HLS and SW/HW partitioning, we are able to easily configure our modular multiplier with large parameters (larger than those seen in the literature) suited for high security requirements. In addition, our modular polynomial reducer can be defined as any generic (cyclotomic) polynomial, allowing optimizations in the homomorphic context. We demonstrate the efficiency of the approach on many designs of full modular

*Asma Mkhinini, Univ. Grenoble Alpes, CNRS, Grenoble INP[**], TIMA, F-38000 Grenoble, France, asma.mkhinini@univ-grenoble-alpes.fr Institute of Engineering Univ. Grenoble Alpes

polynomial multipliers satisfying different area/latency trade-offs. So, our results can guide a designer in his choice of the appropriated configuration with respect to the targeted application.

The paper is organized as follows. In section 2, the background information is introduced. Section 3 presents the related works. Section 4 describes our proposed design. Implementation details are reported and discussed in section 5.

2. Theoretical background

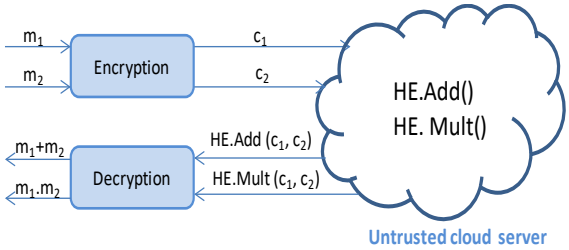
2.1. Homomorphic encryption

The purpose of homomorphic encryption is allowing computations on encrypted data. This means that if a user has a function called f and wants to get $f(m_1, m_2, \dots, m_t)$ for some plaintext messages $(m_1, m_2, \dots, m_t)$, it is possible to instead compute on the corresponding ciphertexts$(c_1, c_2, \dots, c_t)$obtaining a result which decrypts to $f(m_1, m_2, \dots, m_t)$.

Formally, if $\text{Encrypt}(m_i) = c_i (i = 1..t)$ then when evaluating a function f homomorphically on$(c_1, c_2, \dots, c_t)$, we get:

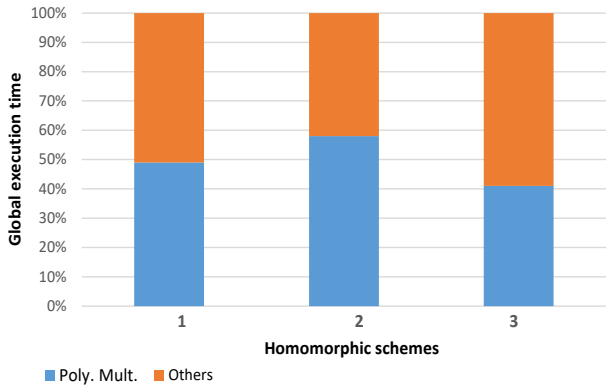
$$\text{Decrypt}\,[f(c_1, \dots, c_t)] = f(m_1, \dots, m_t)$$

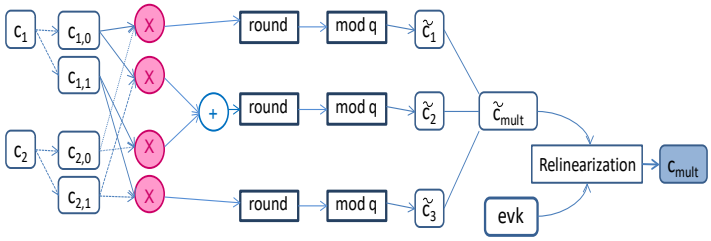
Since every function f can be expressed as a series of additions and multiplications over some algebraic structure, a homomorphic encryption scheme can be defined as an augmented encryption scheme with two additional functions HE.Add() and HE.Mult() to add or multiply on ciphertexts, that result in a ciphertext encrypting respectively the sum or the product of the underlying plaintexts. Figure 1 shows an example of application of FHE in the context of cloud computing. When the number of successive additions HE.Add() and multiplications HE.Mult() can be unlimited during the evaluation step (i.e., computation of the function f), the scheme is known as a fully homomorphic encryption scheme. This generally requires to periodically refresh the ciphertext, otherwise it will be impossible to decrypt. This operation is performed in the encrypted domain and is called "bootstrapping".



2.2. Analysis of software implementations of homomorphic schemes

In order to define the most frequent and time consuming functions during the execution of homomorphic encryption schemes, we profiled existing software implementations [6] [8] of three RLWE-based cryptosystems. Figure 2 shows the profiling results. The analysis reports that polynomial multiplication consumes 41% to 58% of the total execution time. The polynomial multiplication is needed in the encryption, decryption, and evaluation (homomorphic multiplication) steps, as well. For instance, we resume in Figure 3 the operation flow of homomorphic multiplication of two ciphertexts (c_1, c_2) in case of scheme 2 (see Figure 2). In practice, we need 4 polynomial multiplications in order to get c_{mult}. For this reason, a hardware acceleration of this function is of great interest. Hence, an optimized implementation of the modular polynomial multiplication is the target of this paper.





2.3. Modular polynomial multiplication in RLWE-based schemes

In RLWE-based cryptosystems, the primitives are defined over a modular polynomial ring of the form $R_q = \mathbb{Z}_q[x]/f(x)$ where f(x) is a specific irreducible polynomial (cyclotomic polynomial) of degree n and q is an integer modulus ($q > 0$). Parameters n and q define respectively the degree and the coefficient size of polynomials in R_q. Operating in R_q requires reductions modulo q and modulo f(x).

Let $A(x) = a_{n-1}x^{n-1} + \dots + a_0 \equiv (a_{n-1}, \dots, a_0)$ and $B(x) = b_{n-1}x^{n-1} + \dots + b_0 \equiv (b_{n-1}, \dots, b_0)$ be two polynomials of R_q. Computing $C(x) = A(x) \times B(x)$ in R_q needs to first compute the polynomial multiplication of A(x) and B(x) and then reduce the result modulo $(q, f(x))$.

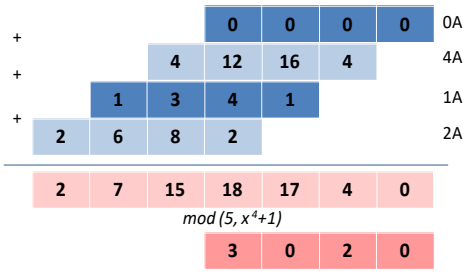
Here is a simple example where $n = 4$, $q = 5$ and $f(x) = x^4 + 1$. We choose:
$A(x) = x^3 + 3x^2 + 4x^1 + 1 \equiv (1,3,4,1)$ and
$B(x) = 2x^3 + 1x^2 + 4x^1 + 0 \equiv (2,1,4,0)$
As shown in figure 4, the result $C(x) = A(x) \times B(x)$ mod $(q, f(x))$ is equal to C_{inR} (3,0,2,0).

2.4. Parameters Derivation

The schemes based on the RLWE problem are governed by a number of inter-related parameters. The modulus q and the degree n are chosen in order to satisfy a given security level λ and a given multiplicative depth L (defined as the maximal number of multiplications that the scheme can handle before it becomes

necessary to apply the bootstrapping procedure). The derivation of these parameters is getting increasing attention lately, in order to provide easy-to-use guidelines for real world applications. In a nutshell, the methodology for parameters extraction aims at sizing these parameters in order to respond to the desired trade-off between security, efficiency, and correctness. Real use-cases of homomorphic cryptosystems define requirements for the multiplicative depth L and the security level λ, then one needs to choose the corresponding values of n and q. Figure 5 illustrates the wide space of practical parameters for RLWE-based schemes with different constraints on L. These configurations are extracted from [6] where the authors explain how to choose these parameters in order to guarantee correctness and security against lattice attacks. They use a lattice basis reduction algorithm based on the van de Pol and Smart approach. This algorithm determines an upper bound on the modulus in a given dimension and for targeted numbers of multiplicative depth L, to ensure a given security level.

For example, with L set to 1, polynomials with a degree around 1024 and coefficients on less than 100 bits can be sufficient. But another scheme requires at least a degree $n = 101853$ and a coefficients size $\log_2(q) = 278$ bits to achieve a security level $\lambda = 80$ bits and a multiplicative depth of 20. Consequently, if we want to target a large set of real applications, our design must be flexible and accept such variations of the parameters.

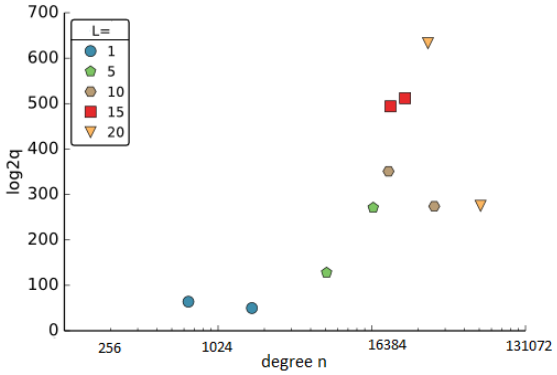


3. Related works

Two principal ways are employed in order to accelerate homomorphic cryptosystems: hardware implementation and GPU (Graphic Processing Unit) acceleration. Hardware accelerators focus mainly on accelerating the most complex functions of homomorphic encryption schemes. There has been some research already conducted into hardware implementations of RLWE-based schemes and their related building blocks. Almost all of them focus on the polynomial multiplication.

Many of these implementations have used the Number Theoretic Transform (NTT) or the Negative Wrapped Convolution (NWC) to perform polynomial multiplication in R_q efficiently. NTT and NWC are two special forms of the Fast Fourier Transform (FFT), known as the asymptotically fastest algorithm for computing polynomial multiplication.

In [10], Doröz et al. propose an implementation of the modular polynomial multiplication computed with the NTT algorithm and a Barrett reducer. A pre-computation based on the Chinese Remainder Theorem (CRT) is performed on input polynomials to reduce the size of coefficients. The overall architecture is based on an array of units, which gives some flexibility to process several residue polynomials in parallel. They evaluate their architecture on polynomials of fixed degree $n = 2^{15}$, and fixed coefficients size $log_2(q) = 32$ bits. Their accelerator was dedicated for a specific homomorphic scheme. In [11], Chen et al. present an optimized design of the modular polynomial multiplication.



All computations are carried out in the FFT domain on polynomials with degree n ∈ [256, 2048] and coefficients size ∈[20, 29]. They provide a selection method for the parameter set supporting efficient modular reduction, meeting at the same time the security requirements for RLWE and Somewhat Homomorphic Encryption (SHE) schemes. SHE means that the maximum number of successive operations in the encrypted domain is limited. Though efficient, this selection leads to many restrictions on the polynomials supported by the design: polynomial modular reduction is in fact computed with respect to the common choice $f(x) = x^n + 1$. When FFT multiplication using NWC is employed, they show that the modular reduction $(x^n + 1)$ is eliminated; this elimination comes at the expense of pre- and post-computation steps. A hardware architecture for the modular polynomial multiplication is described in [13]. They provide a fast unit for polynomial operations using CRT and NTT for multiplication combined with an optimized memory access scheme and a Barrett reduction method. The implemented unit can be used to instantiate a specific encryption scheme. Results are provided for $n = 32768$ and $log_2(q) = 1228$ bits. The authors of [14] use the Karatsuba algorithm to implement the modular polynomial multiplication in hardware. They demonstrate that for various degrees and coefficient sizes, Karatsuba can be a good alternative to FFT. Lastly, Jayet-Griffon et al. [12] consider the polynomial multiplication of 512-degree polynomials with 32-bit coefficients. They analyze and compare three algorithms (Karatsuba, FFT, and Schoolbook): the Schoolbook method is shown as the most efficient for a hardware implementation, due to its simple and regular structure. Modular reductions were not covered in their work.

All the aforementioned related works reported significant speed up factors when compared with software implementations. This speed up illustrates that further research into hardware implementations could greatly improve the performance of FHE schemes. However, they are almost all designed for fixed length operands and optimized for one specific type of multiplication algorithm (mainly FFT-based algorithms) which puts restrictions on the parameters selection. This makes them limited to some specific schemes and application domains. Besides, the operand

sizes are not very large, which has a direct impact on the security level of the cryptosystem. When modular polynomial reductions are performed, the simplest (and more limiting) choice is often selected.

Consequently, existing designs can be considered as proof-of-concept implementations only suitable for homomorphic encryption with small multiplicative depth circuits and low computation complexity. In the homomorphic context and because of its rapid growth evolution, new designs with a wide range of parameters are needed. These parameters are very large, consume large amounts of memory, and require many resources in order to perform efficient computations. Thus, memory storage and available resources in a target device should be taken into consideration, especially when manufacturing a specific circuit is not affordable and FPGA (Field Programmable Gate Array) platforms are therefore intended.

4. Software/Hardware design description

4.1. General presentation

We propose a hybrid and flexible SW/HW design based on a generic polynomial multiplier. This design is basically constructed from two parts: a dedicated hardware accelerator, and the software running on a general-purpose processor. Our solution aims at improving the overall performance and supporting much larger parameter sets than previous designs while optimizing resources for a given computation performance level. As in most previously published approaches, we will demonstrate our solution on FPGA-based implementations. However, the same approach may also be used with pre-characterized libraries to generate an application-specific integrated circuit (ASIC). The general-purpose processor can be implemented on-chip, or the hardware part of the accelerator can be connected to a computer. The interface between hardware and software can thus be performed, for example, through a high performance AXI (Advanced eXtensible Interface) bus, when software is running on an embedded processor, or a PCI (Peripheral Component Interconnect) express bus in the case of a computer processor. The choice should take into account the performance targeted for the global design, the communication and post-processing overheads, as well as the implementation constraints.

4.2. Design configurability

Our accelerator has been designed to support polynomials of any degree and any coefficients size. This goal has been met thanks to four hierarchy levels. As shown in Figure 6, level 1 and level 3 compute the product of two polynomials with large degree, while levels 2 and 4 deal with the product of coefficients with large sizes. This approach allows us to design efficient implementations for the lower (smaller) blocks, and configurable algorithms for the upper (larger) ones.

The designer starts by defining the input parameters (degree, coefficients size and irreducible polynomial used for the modular reduction).

In order to compute the multiplication of large input polynomials, we first represent the inputs as sets of polynomials of smaller fixed degree K. Then, we compute the pairwise products of each pair of sub-polynomials, using a hardware block for the multiplication of polynomials of fixed degree K based on the Schoolbook algorithm. This algorithm is not the fastest from a pure algorithmic point of view (asymptotic complexity), but it has three main advantages: it is more efficiently implemented on FPGA targets [12], it does not require pre- or post-processing and it does not impose any specific constraints thus allowing the possibility of optimizations such as batching (see section 5.3). An additional reason for choosing this algorithm is that the proposed decomposition limits the degree of the polynomials at level 3, so the degrees required to take full advantage of the asymptotic complexity of other algorithms is not reached.

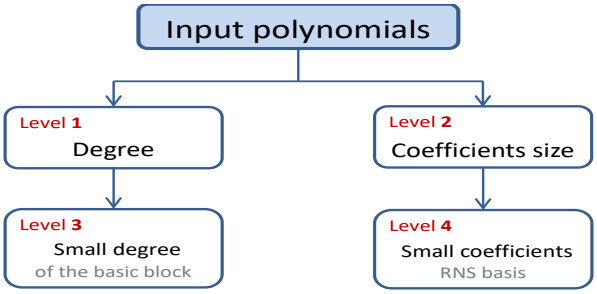


Figure 6. Multi-level design

The product of large coefficients is calculated through a RNS (Residue Number System) approach [15]. The advantage of RNS is that computations can be performed in parallel, that can result in a significant speed-up. We convert each large coefficient into a set of several values of smaller size by applying the RNS transformation. This technique can be easily made (almost) independent of the coefficient size: in order to support larger coefficients, it is sufficient to add a new element to the existing RNS basis without need to change the underlying architecture. A more detailed description of our approach is discussed in [16].

Let us mention that for small degrees and/or small coefficients sizes, level 1 and/or level 2 are optional (see Figure 6), as the algorithm can operate directly on full size operands.

4.3. SW/HW Partitioning

Our hybrid design implements part of the computations in software and part of them in hardware. The adopted partitioning is summarized in Figure 7. On the software-level, we consider simple and cheap computations such as input representations in subset polynomials, RNS basis generation, and polynomial additions; additionally, polynomial modular reduction is also computed on the main processor, in order to take advantage of the flexibility of a software implementation.

The hardware part implements the product of large coefficients in the RNS domain, and the Schoolbook multiplication algorithm at the polynomial level.

In order to reach our goals in terms of flexibility (size of parameters, but also implementation target), the hardware blocks are generated with a High Level Synthesis (HLS) tool.

4.4. High Level Synthesis

High Level Synthesis aims at transforming a generic input algorithm into a Register Transfer Level (RTL) architecture for a given target technology. This allows obtaining better productivity

when compared to classical implementation approaches using direct RTL design in languages such as VHDL or Verilog, as the designer can work on a higher level language which is much easier to maintain.

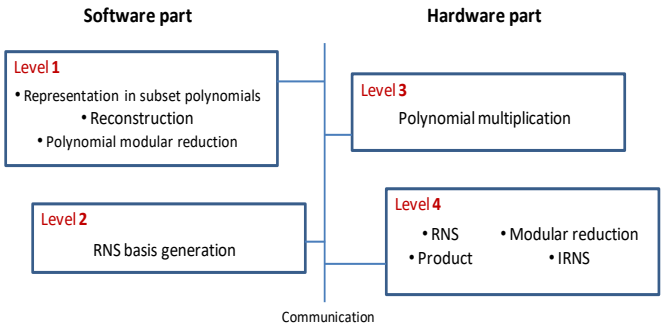


Figure 7. Modular polynomial multiplier: SW/HW partitioning overview

Several HLS tools exist; in this work, we use AUGH [17] since it is open, it may support different targets through the corresponding tool chain, and it can provide early estimations about the performance of the final result that can guide the designer in rapid refinement loops. AUGH is an autonomous HLS tool: it generates RTL descriptions quickly, under only global resource and frequency constraints [18]. This is achieved by performing incremental transformations of the input design description. The small complexity of the design space exploration (DSE) algorithm and the efficient use of all internal circuit structure constraints make this HLS tool very fast and able to generate pertinent solutions.

In this work, we will target two different FPGA technologies and a SoC (System on Chip) from Xilinx, but the methodology can be applied to other targets as well, provided that the corresponding flow is available. In our context, specifying new polynomials just implies to modify the input algorithms described in C language and let the HLS tool produce the RTL descriptions for every new specification (see Figure 8). Changing other parameters is fast and simple as well, as the designer can modify directly the high level description of the algorithm. Similar interventions on a RTL description of the design would take much more time and would be much more prone to errors due to a more complex description code.

AUGH provides different techniques for design optimizations, including unrolling and pipelining for the loops, wiring for the branch conditions and using maximum operator sharing [18]. Each RTL generation is analyzed by the user who can then command the HLS tool (with these directives) to converge towards a better solution in the next trial. DSE process detects possible transformations of the design that bring more parallelism, and applies these transformations until the user resource constraints are reached. Then, the RTL design description is generated.

5. Implementations, comparisons and discussions

5.1. Comparison with the state of the art

In order to provide a fair comparison of our results with the state of the art, we configure our design with parameters as close as possible to [10], [11] and [12] and perform the synthesis on similar targets.

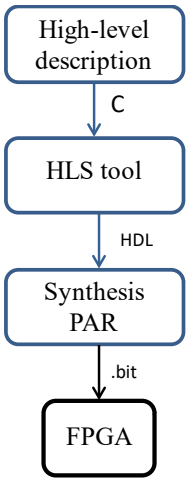


Figure 8. HLS based approach: tool chain

We set parameters for multiplication of polynomials of degree 512, 1024 and 32768 and with coefficients sizes of 26 and 32 bits. Under these configurations and for the smallest degrees, we do not divide inputs into smaller polynomials and we feed our basic block of level 3 directly with the corresponding polynomials.

Table 1 demonstrates that our approach, though about twice slower compared to manually optimized designs, consumes much fewer resources than the two designs reported in [11]. We obtain a reduction by a factor of 9 on average. This significant difference can be explained by the choice of the polynomial multiplication algorithm. For the FTT- and NTT-based algorithms (as in [11]), the pre- and post-computation steps are complex and require storing additional parameters. For the Schoolbook method, no pre- and post-computations are needed and we only have to store the input and output coefficients. Thus, FTT and NTT implementations have lower latency than our accelerator thanks to their lower complexity and their hand-crafted design but have several constraints and require a large amount of hardware resources.

For comparison with [12], where the authors implement the Schoolbook algorithm, we present two solutions in Table 2. In order to exploit parallelism and to achieve the minimal latency, we apply an optimization directive of AUGH not allowing operator sharing. Doing so, we instruct the HLS tool to use as many DSPs as possible: in this case only LUTs are used to store the coefficients and a maximum number of DSPs can be easily parallelized (first design in Table 2). In the other solution (second design in Table 2) BRAMs are used to store information, which limits the number of DSPs that can be used efficiently. It must be stressed, however, that our results are after placement and routing, while [12] only gives results after synthesis. Depending on the choice made with the HLS tool, the latency may be only slightly augmented (+28%) compared with the noticeable gain in resources (more than 35x). Resources can be further reduced up to a factor of 41 with the second solution, but leading to a loss of performance by a factor of 12.6.

For polynomials with degree 32768 used in [10], we divide the inputs into polynomials of degree 8192 and we apply our

approach first without any optimization. In this case (first design in Table 3), the hardware resources required by our design are nearly 28 times less than [10] on average; on the other hand, our design has a latency of 41 ms, while the authors of [10] report a latency of about 9.5 ms. Since our accelerator requires relatively few hardware resources, we can compute in parallel on 4 instances: with this optimization, the second design in Table 3 has a latency comparable to [10], while still maintaining a large advantage in resources.

Globally, these examples of results for our designs show that they consume fewer resources than the state of the art thanks to the proposed approach, while achieved performances can be close to hand-made designs. Using fewer resources also leads to possible parallelization of several instances while meeting the resource constraints of a given FPGA; in that case, even better performances can be reached. The approach offers at the same time a large degree of configurability, allowing the designer to change several important parameters on-the-fly. In addition, our approach is highly flexible, since the same generic high-level description can be used to produce a new circuit with different area/performance trade-offs i.e., we can generate a very cheap (but slow) polynomial multiplier, or a fast but more resource consuming one. Other possible optimizations can be applied, as it will be shown in the next sections.

5.1. Design space exploration

With our flexible design, we can instantiate our architecture with different coefficients size and different degrees, targeting several platforms. Larger parameters imply large values of n and q that grant high security levels and significant multiplicative depths.

When we handle large coefficients, our approach proposes to transform them into their respective RNS representation. The choice of the RNS basis (size and co-prime modulus) is primarily based on the coefficients size of the input polynomials and the available resources on the target device. The key idea is to take advantage of the parallelism offered by the RNS representation and speed up our computations. If we ever need to increase the supported size, it is sufficient to extend the RNS basis without need to change the underlying architecture.

The RNS basis size and modulus are kept as parameters as well as the degree K of the basic block performing the Schoolbook multiplication. Let us now fix for example the coefficients size at 64 bits and vary the degree from 8192 to 32768 to cover a new range of parameters. For each degree, we choose a different degree K of the basic block. When the resources of our target are sufficient, we perform computations in parallel on several instances.

Figure 9 shows that some configurations are more efficient than others. In fact, when computations are running on P instances in parallel, we roughly multiply the resources by P and divide the latency by P (illustrated by the cases of n=8192 and n=32768, figure 9). But, setting for example K=1024 to calculate the multiplication of polynomials with a degree 16384 is not an appropriate configuration. The performance loss is essentially due to 16^2 calls to the basic block *Polynomial_K*. When changing K and/or applying an optimization, the latency is also affected by the cost of the data transfers and therefore some configurations are not suitable.

Table 1. Comparison on Spartan 6 with polynomials of degree 1024 and 26-bit coefficients

References	**Resources**				
	Slice LUT	Slice Register	DSP	BRAM	Latency (μs)
[11](1)	10801	3176	0	0	40.98
[11](2)	2464	915	16	14	32.28
Our work	182	114	3	10	69.1

(1) Multipliers are built by pure LUTs (2) Use DSPs and Brams

Table 2.Comparison on Virtex 7 with polynomials of degree 512 and 32-bit coefficients

References	**Resources**				
	Slice LUT	Slice Register	DSP	BRAM	Latency (μs)
[12]*	252341	130826	512	2048	4.11
Our work (1)	7032	920	368	0	5.27
Our work (2)	171	102	3	3	66.41

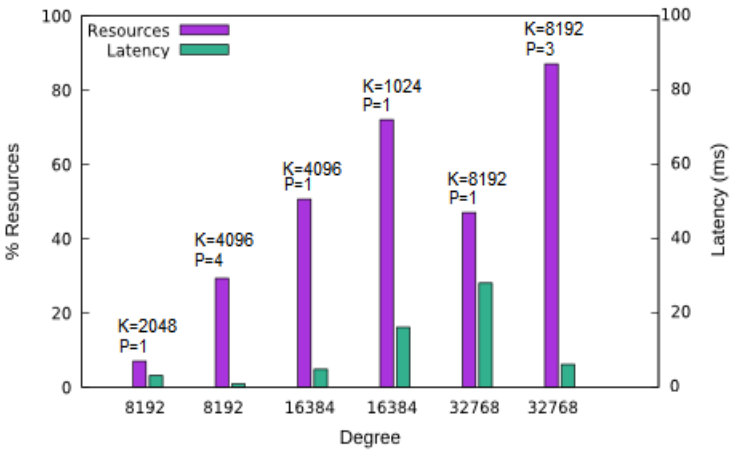
(1) Data is stored in Lutrams (2) Data is stored in Brams *Synthesis results

Table 3.Comparison on Virtex7 with polynomials of degree 32786 and 32-bit coefficients

References	**Resources**				
	Slice LUT	Slice Register	DSP	BRAM	Latency (μs)
[10]	219192	90789	139	768	9.51×10^3
Our work (1)	3392	1920	48	792	41.12×10^3
Our work (2)	13568	7680	192	792	10.28×10^3

(1) Our approach without optimizations (2) Several instances of the basic block running in parallel

Figure 9. Performance of our designs on polynomials with 64-bit coefficients and



different configurations (degree K, parallelization, degree P)

Several multipliers with different area/performance ratios can be generated by modifying the RNS basis and the degree K of the basic block, and also by the parallelization of the computations on many instances of the basic block. Thanks to the HLS design flow, rapid feedback on circuit characteristics is used to evaluate deep architectural changes in short time and pick up the more suitable parameter sets. Then, the designer can select the design satisfying his constraints among the set of generated circuits. The development timeline of a new solution is about 3 hours on average, which is a very short time in comparison with a hand-made design.

5.2. Our system performances

In this section, we evaluate the performance of our hybrid modular polynomial multiplier with respect to the communication times between hardware and software. We run computations on polynomials of degree 512 with 32-bit coefficients and we pick two choices for the modular polynomial:$f(x) = x^{512} + 1$, and a general cyclotomic polynomial of degree 512. This choice is motivated by the possibility, allowed by a generic polynomial, of applying the batching optimization [19].This technique can be used in order to pack multiple messages into one single ciphertext, therefore allowing parallel homomorphic evaluations. Hence, it permits great versatility in the computations and improves performance. This technique is based on the CRT theorem and requires that the polynomial f(x) is different from $x^n + 1$.

To transfer data, we consider an AXI interconnection between the hardware and software part. Our platform is the Zybo Zynq-7000. In our case, we send to the FPGA two input polynomials of 32*512 bits each and we receive one polynomial of 73*1024 bits. We use one high performance AXI bus to send the input coefficients and two high performance AXI buses to receive the output polynomials: this is due to the fact that the polynomial reduction is made in software, and the multiplication result is thus (almost) twice larger than the operands. With this solution, we can enhance the performance of our system and have the final result faster: a speed up factor of about 2.98 is obtained compared with a solution with only one bus, as the hardware acceleration is mitigated by the data transfers.

Our architecture reports a global latency, including transfers, of 0.75 ms when $f(x) = x^{512} + 1$ and of 1.89 ms when f(x) has a general form, which is still very fast compared to pure software implementations.

When compared with the state of the art, we have shown that our implementations consume few resources but report smaller speed-up factors. To counter this, we have proposed to run computations on many instances of the basic block in parallel. Figure 10 illustrates an optimized proposition when performing computations on 4 instances of blocks with degree 512 in parallel.

5.3. System evaluation on larger parameters

In this section, we evaluate our approach on designs with larger parameters. For each configuration, we choose two forms of the irreducible polynomial f(x). The first one corresponds to the popular choice $x^n + 1$. The second one has a general form and allows optimizations in the homomorphic context. Tables 4 and 5 provide implementation results and comparisons with software implementations we developed with Sage and run on a Intel Core i5-2450M (2.5 GHz). We decided to make such a comparison in order to get the same parameters configuration of parameters and because published works do not cover such parameters. Our reference embedded platform (the Zybo board) has processing power and memory that are not even comparable to desktop or server CPUs. Nonetheless, and withstanding the overhead of the data transfers through the AXI bus, the results obtained are quite interesting.

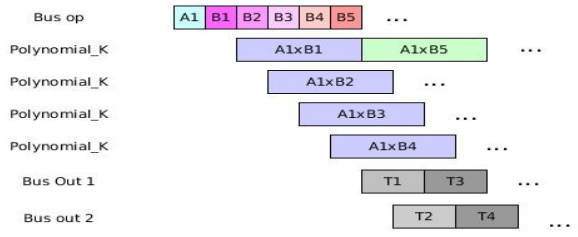


Figure 10. Parallelization of the computations on many instances of the basic block *Polynomial_K* [1].

Table 4. Implementation results (ms) when $f(x) = x^n + 1$

n	q	Software implementation	Our design
4096	32	32.6	2.1
8192	64	76.7	5.8
16384	128	162.8	11.7

Table 5. Implementation results (ms) when f(x) has a general form

n	q	Software implementation	Our design
4096	32	41.8	4.1
8192	64	88.5	12.7
16384	128	202.9	21.8

Results show that our approach reports significant speed-up factors when compared with pure software implementations, in spite of the large parameter values. The latencies when the

polynomial reduction is computed modulo $x^n + 1$ are better than the reduction modulo a polynomial with general form. This is due to the complex form of the polynomial in the second case (only 2 coefficients vs. (n+1) coefficients) and to the number of iterations required to get the final results. For a fair comparison, it must be stated that such a general form is not covered by most works in the literature, especially when hardware implementations are considered.

Several software implementations and libraries dedicated to homomorphic cryptography exist today [20], **Error! Reference source not found.**[21], which achieve very interesting performance when executing on high-end processors. These implementations are usually based on the NTT-based algorithm in order to speed up the modular multiplication on very large operands. This algorithm, however, has very large requirements in terms of memory usage, it constrains the choice of parameters, and it is usually optimized by exploiting the advanced instruction set available on modern processors, such as SSE and AVX. When targeting limited devices and/or current client-server frameworks, those implementations cannot be used directly because of the memory cost, or do not perform well enough due to missing advanced instructions. For this reason, a comparison with these works is difficult, since different application domains and platforms are targeted.

6. Conclusion and future works

We present a flexible and generic SW/HW co-design for the modular polynomial multiplication, the most computationally intensive operation in homomorphic cryptosystems based on the RLWE problem. Given the large parameters required in such schemes, we propose an RNS based algorithm and an efficient decomposition of the large input polynomials. Our design can be easily configured thanks to a HLS approach and sets no restrictions on the parameters that define the RLWE problem leading to high security levels and large multiplicative depths when necessary. Our approach can also optimize the accelerator for applications requiring small parameters. Our architecture can be instantiated to accelerate any RLWE-based scheme; additionally, even if the proposed methodology has been illustrated only on the polynomial multiplication, it can be used to implement and accelerate other primitives required by homomorphic schemes.

Future works include more complex communication schemes, such as using two AXI High Performance input ports and thus increase the number of instances performing in parallel. As the memory is one major bottleneck, we can reduce the amount of memory by controlling and scheduling the loading of sub-polynomials $A_i(x)$ and $B_j(x)$. It may also be interesting to evaluate the benefits of a hardware implementation of the polynomial reduction. Other primitives used in the FHE schemes may also be implemented in hardware using the same methodology. Finally, these accelerators will be integrated to evaluate a full homomorphic scheme on an FPGA/processor platform targeting embedded applications, and the performance will be compared to state-of-art software libraries ported to the same environment.

References

[1] A. Mkhinini, P. Maistri, R. Leveugle and R. Tourki, HLS design of a hardware accelerator for Homomorphic Encryption, IEEE 20th International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS'17), pp. 178-183, Dresden, Allemagne, 19 au 21 avril 2017.

[2] C. Gentry, A fully homomorphic encryption scheme, Ph.D. dissertation Stanford University, 2009.

[3] Joppe W. Bos, Kristin Lauter, Jake Loftus, and Michael Naehrig. Improved security for a ring-based fully homomorphic encryption scheme. In Martijn Stam, editor, Cryptography and Coding, volume 8308 of LNCS, pages 45–64. Springer Berlin Heidelberg, 2013.

[4] Junfeng Fan and Frederik Vercauteren. Somewhat practical fully homomorphic encryption. Cryptology ePrint Archive, Report 2012/144, 2012. http://eprint.iacr.org/.

[5] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. In ITCS 2012, pages 309–325, Cambridge, Massachusetts, 2012. ACM.

[6] Tancrède Lepoint and Michael Naehrig. A comparison of the homomorphic encryption schemes FV and YASHE. In Progress in Cryptology - AFRICACRYPT 2014,7th International Conference on Cryptology in Africa, Marrakesh, Morocco, May 28-30, 2014. Proceedings, pages 318-335, 2014.

[7] Zvika Brakerski. Fully homomorphic encryption without modulus switching from classical GapSVP. In Advances in Cryptology - Crypto 2012, volume 7417 of LNCS, pages 868–886. Springer, 2012.

[8] Shai Halevi and Victor Shoup. Algorithms in helib. In Advances in Cryptology-CRYPTO 2014 - 34th Annual Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2014, Proceedings, Part I, pages 554-571, 2014.

[9] Kim Laine and Rachel Player. Simple encrypted arithmetic library - seal (v2.0). Technical report, Microsoft Research, September 2016.

[10] Y. Doröz, E. Öztürk, E. Savas, B. Sunar , "Accelerating LTV Based Homomorphic Encryption in Reconfigurable Hardware",CHES '15, 185-204.

[11] D. D. Chen, N. Mentens, F. Vercauteren, S. S. Roy, R. C. C. Cheung, D. Pao, and I. Verbauwhede, "High-speed polynomial multiplication architecture for ring-lwe and SHE cryptosystems," IEEE Trans. on Circuits and Systems, vol. 62-I, no. 1, pp. 157–166, 2015.

[12] Cedric Jayet-Griffon, M.-A. Cornelie, P. Maistri, Ph. Elbaz-Vincent and R. Leveugle. Polynomial Multipliers for Fully Homomorphic Encryption on FPGA. In 2015 International Conference on ReConFigurable Computing and FPGAs (ReConFig), Mexico City, 2015, pages 1-6.

[13] S. Sinha Roy, K. Järvinen, F. Vercauteren, V. Dimitrov, and I. Verbauwhede, Modular hardware architecture for somewhat homomorphic function evaluation, in Proc. of Cryptographic Hardware and Embedded Systems – CHES 2015. Springer, pp. 164–184.

[14] V. Migliore; M. Mendez Real, V. Lapotre, A. Tisserand, C. Fontaine, G. Gogniat. Hardware/Software co-Design of an Accelerator for FV Homomorphic Encryption Scheme using Karatsuba Algorithm. In IEEE Transactions on Computers, vol.PP, no.99, pp.1-1.

[15] H. L. Garner, The residue number system, IRE Transactions on Electronic Computers, vol. EC-8, no. 2, pp. 140–147, Jun. 1959.

[16] A. Mkhinini, P. Maistri, R. Leveugle, R.Tourki and M. Machhout, A Flexible RNS based Large Polynomial Multiplier for Fully Homomorphic Encryption, 11th IEEE International Design & Test Symposium (IDT), Hammamet, Tunisia, December 18-20, 2016, pp. 131-136.

[17] A. Prost-Boucle, "Augh project" 2016, [Online]. Available: http://tima.imag.fr/sls/research-projects/augh/

[18] A. Prost-Boucle, O. Muller, and F. Rousseau, Fast and standalone design space exploration for high-level synthesis under resource constraints, Journal of Systems Architecture, vol. 60, n. 1, 79–93, 2014.

[19] N. P. Smart, F. Vercauteren, Fully homomorphic SIMD operations, Des. Codes Cryptography 71(1): 57-81 (2014).

[20] Seal : Simple Encrypted Arithmetic Library, https://sealcrypto.codeplex.com/

[21] Helib library, https://github.com/shaih/HElib

[22] NFLlib library, https://github.com/CryptoExperts/FV-NFLlib

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 434-442 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

A Smart Mobile Application for Assisting Parents in Anti-Drug Support

Tsz Hei Yeung, Chi Kit Ng, Vincent Ng*

Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Article history:
Received: 01 December, 2017
Accepted: 15 January, 2018
Online: 10 February, 2018

Keywords :
Drug Abuse
Mobile Application
Location Check-in
Behaviour analysis
Instant Messaging

ABSTRACT

Drug abuse problem is not a new problem, but it is still serious in Hong Kong. The emergence of hidden drug abuse youth has brought more difficulties to them and their parents. The usual ways of providing anti-drug abuse assistance are not sufficient to these types of young people. Intelligent methods have been developed to assist anti-drug workers. In this paper, we present three mobile applications with a set of intelligent tools for parents and social workers to help our young people. The applications have an alerting tool to detect the possibility of drug abuse of a young student, and to alert corresponding parents and social workers for remedial actions. Another function is the nearby notification tool to help social workers in identifying people who may need assistance. For the alerting function, a number of data mining and text analysis techniques have been adopted for behavior analysis. A supporting instant messaging has been co-developed for communicating alerting messages.

1. Introduction

Drug abuse problem is not a new problem, but it is still a big problem in Hong Kong Society. Even there is a tendency of decreasing in drug abusing, the government report indicated that the age of first drug taking is decreasing. This shows that the problem is in serious with depth (age) but not the width(quantity). Besides, from the perspective of teenagers, they do not have much knowledge on how to get the treatment service from different organizations.

Furthermore, Hong Kong has another problem which is the hidden youths or "socially withdrawal". The term "hidden youths" is originated from Japanese word "Hikikomori" and it means that a person is scheduled at home by at least six months [1]. There is one more level under the "hidden youths" which is "Otaku" in terms of Japanese. There are different characteristics for "Otaku", but the commonly use is that "Otaku" always stay at home for activities and avoid social activities. A study found that these people are satisfied with their quality of life [2]. Also, it is difficult for other people to interact with them normally. And once they are drug abuser, they become the riskiest group of people as people cannot discover them easily and help them return to life. In Hong Kong, it has been reported that, since 2010, 80% young drug abusers take drugs at home.

*Vincent Ng, Department of Computing, The Hong Kong Polytechnic University, 852-2766-7242, Email: cstyng@comp.polyu.edu.hk

According to Narcotics Division, the median drug abuse history of newly reported abusers (i.e. the time for abusers to be reported to the CRDA by reporting agencies from their first drug abuse) remained between 1.7 and 2.1 years during 2007 to 2009. It then increased continuously to 5.9 years in 2015, and dropped to 4.6 years in 2016 [3]. One reason is because psychotropic substance abuse is more popular among young drug abusers [4]. Unlike traditional drugs such as heroin, psychotropic substance abuse is more "hidden" in nature. Young drug abusers tend to take psychotropic substance at home. It is hard for us to bring the young drug abusers to light.

Smartphones have become popular among young people. There are close to 5 million Hong Kong persons aged 10 and over had smartphones in 2014 and the smartphone penetration rate has exceeded 77.2% [5]. Developing anti-drug mobile applications can be an effective way to pass healthy information to young people as well as parents and social workers. Hence, in 2014, we submitted a proposal to the Beat Drugs Fund in HKSAR to develop a series of mobile applications for students, parents and social workers with a backend platform to support information sharing.

In our design, we aim to find out the hidden or potential drug abusers, and to warn their family and anti-drug workers about the problem. Another important function is to support the nearby function for social workers to access young people. For the rest of the paper, Section 2 provides the background review. Next, in Section 3 and 4, we describe our application design and

implementation approach. Section 5 concludes our present work in this paper.

2. Background Review

2.1. Anti-Drugs and Behavior Analysis Mobile Apps

Dealing with the drug abuse problem, different parties in Hong Kong society have developed different mobile applications to link with the youths. There are two mobile applications recently developed in HK and they are “Anti-Drug Information Center” and “毒癮無可忍” to deal with the drug abuse problem. The “Anti-drug Information Center” (禁毒資訊站) developed in Macau provides the drug information as shown in Figure 1 and can support parents as a simple drug abuse detection tool [6].

Figure 1. Home screen of the Anti-drug Information Center

This mobile application adopts an expert system approach which is based on the anti-drug resource kit designed by the T.W.G.Hs. CROSS Centre [7]. However, there were criticisms on the test because the behavior criteria could have different interpretations, not only for detecting drug abuse issues. For example, it simply identifies “emotionally unstable” and “performance mysterious contact with friends” as symptoms of drug abuse, but these symptoms can also be characterized adolescent.

For the ”毒癮無可忍” application, it was developed by a Christian organization. It also provides the drug related information and detection test. The information it provided are the real cases sharing, information of referral center and some encouragement message from the bible. The detection test is rule-based and has a number of questions as shown in Figure 2. That is, when your answer meets all criteria of detection, response will then show up. The responses are video formats so as to arise interest and attention of the young people.

Drugout is an iOS Apps aims to passing knowledge about anti-drug [8]. In the application, players will try to help a game avatar to fight with his drug abuse issues as shown in Figure 3. Score will be accumulated during the actions, and players can get bonus score if they invite others playing the application. Finally, there will be real prizes given to the high score players. Overall, it is an interesting game that can attract students to play and be able to deliver the anti-drug information to the players. The bonus scoring can encourage players to invite more people play the game and players can experience what a drug abuser will suffer from the drugs. Yet, it only provides information without any communication platform. Also, its use is for the youth and parent may not get help from this application.

Figure 2. Addiction testing

Figure 3. Dragout

Another similar mobile application is called Detoxification [9]. Besides anti-drug information, it also provides interesting game and situational questions which raise awareness of drug abuse as shown in Figure 4. Yet, the detoxification application has not been completed and it also lacks of resource links to support youths when they encounter a problem.

Another type of mobile applications is for behavioral analysis, such as the “Pro-sociality Behavior Test”. It contains a self-assessment to test pro-sociality behavior tendency. Its problem is having too many questions for an assessment and users often feel tired when doing it. There is a similar mobile application, “Depression Test”, for depression testing [10]. It provides a nice historical review and users can see the changes of their depression levels. The application provides no connection between users and social service organizations. Users cannot find an easy way to seek for help if they encounter a problem.

Although there are many mobile applications about behavior analysis in the app stores, most of them are not related to anti-drug.

This motivates us to develop a suitable solution with design objectives below:

- Provide an accurate prediction based on behavioral information provided by users.
- Able to detect potential drug abuse with a few guided questions with their answers.
- Able to provide connection and interaction channels between users and anti-drug social service organization.

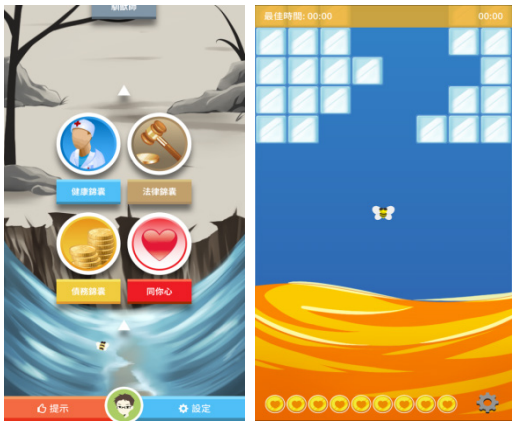


Figure 4. Detoxification

2.2. Alerting Systems

Alerting systems have been applied in different service domains. The Brigham Integrated Computing System (BICS) provided a clinical alerting system at Brigham and Women's Hospital (BWH) since June 1994 [11]. It used an event monitor to determine whether a new clinical data warrants an alert. The notification program would send an automated e-mail message messages to a duty physician, or through other means of communication. The physician could then log on any computer to review the alert and take immediate therapeutic action. The alerting system reduced the time it took for physicians to respond to critical laboratory results. Physicians thought that the communicated information was important and the service can help clinicians to be well informed of important abnormalities in the flood of data they have continually receiving.

Alerting system is also essential for providing the location and information about vehicles to passengers, owners or users [12]. Bus transportation system has been proposed to track a large number of buses simultaneously and detects the routes and directions automatically through GPS satellite [9]. It used a web-based, mobile communication and SMS platform for communicating bus arrival times and predicted the real-time inter-station speed.

Email alerting has been proposed with education, health care and other life applications. They can be used for two-way knowledge translation, which involves receiving case situation to professionals and then sending their evidence-based recommendations. In [13], it is reported that professionals were familiar with email communication. While many felt comfortable and liked using emails in their professional life, yet some others felt overwhelmed by the volume of emails they receive. The study also suggests that email alerting can contain valid and trustworthy information as an easy way to disseminate information to multiple recipients. This would give the reader an option to read emails immediately or later.

Instant messaging alerting is another frequent channel to support alerting messages. The major advantage of instant messaging alerting is real time or instantly delivered. When a receiver is online, the message will be received as soon as the user sends it. Also, instant messaging alerting tends to be confined to a single line and easy to read [14]. However, instant messaging systems are often of proprietary platforms and cross platform support is difficult. Communicating via instant messaging requires that both parties have accounts with the same instant messaging service.

With the above reviews, we have proposed to use multiple channels for sending our altering messages, both email and instant messaging. As social workers are familiar with both channels, this multi-channel platform can effectively work for their clients.

2.3. Location Sharing

Check-in provides location awareness support so that it can act as a channel to share, comment and rate any spots in different areas. It aims to establish a location-based social network (LBSN) for users. All users included students, parents, social workers can publicize their current locations to others. Through this check-in step, users can contact their friends nearby and chat with instant messaging. A special function for social workers is that they can be notified the youths' current locations if youths choose to open the information.

In Google Map, users can leave comments and rate places. The comments and ratings are public to everyone in the world. More than comments and rating, users can also upload and share photos about the spot. Sharing the location with user's friends through other applications like Facebook and Whatsapp, is available in Google Map [15]. A good example is an application developed by Apple Inc., FindMyFriends. It allows users to follow and track friends with their mobile device. Users can also choose to share their location with others. The app determines users' location using GPS in the device. If the user is requested by his friends to see their current location, users can contact their friends on the map immediately with the contacts in their device [16].

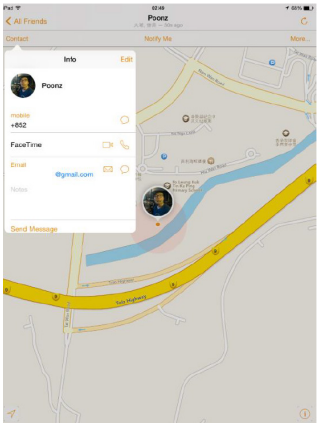


Figure 5. FindMyFriends

Another similar application is Swarm which is developed by Foursquare. It is a companion application of Foursquare. Users are allowed to share their locations within their social network. A check-in can be broadcasted to through other channels like Facebook or Twitter. User can easily see nearby friends, make future plans as well as checking in a location. Figure 6 shows a check-in page of Swarm.

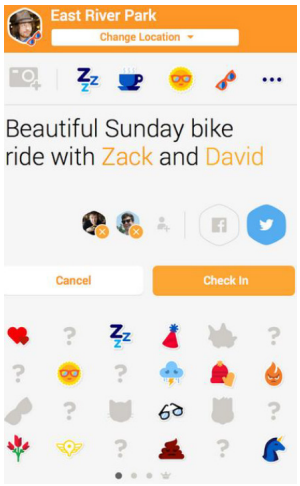


Figure 6. Swarm in Foursquare

Swarm has all the required functions for our objectives in the aspect of location sharing. From "check-in" and "check-out" to chat with friends. However, because our target users require a high level of confidentiality, only secured sharing platform is allowed.

3. Design Methodology

After the review the current problems in the previous sections, a set of mobile application is decided to develop to suit the needs from the different users. For parents, the mainly functions will help them on drug abuse detection of their kids and they can use the application to seek for professional help and build a community between other parents and social workers. For students, the application will try to collect their data that will be useful to the project team, also the application will help the youth have more connection with their parents, friends and social workers by instant messaging or share their location information. For social worker, they should able to use the application to upload the anti-drug information and give help to the youths and their parent. With the types of users, 3 different mobile applications have been designed and developed and they are called Aurora, Giai and Theoi, respectively.

In the rest of this section, the overall design and three functional components are presented. The first component is on the youth behavioral analysis for users to predict a potential drug abusing case. The second component is to support the alerting function when such case is detected. It can effectively work for their clients. The last is about the location sharing and near-by search support.

3.1. Overall Design

At the beginning of our work, teachers, principals, students, social workers and anti-drug experts have been consulted and surveyed on the suitable functions of the three applications. For the student mobile application, it includes the following functions:

a) Head News Center: An information receiving channels for youths to receive anti-drug information, such as events, activities. Some of the news can be in a form of alerting and pop-up such that youths would receive something interesting to consider. It is a way to encourage youths to get more positively involved in the community. An interest profile can be set by the young user for customized information.

b) Q&A Section: a resourceful function for youths to receive answers of common questions, such as where and how they can seek help.

c) Survey (game corner): it is useful to collect youth's current status from time to time. They can provide some of their own data on a voluntarily basis with a bonus point encouragement. The bonus point scheme can be supported by small souvenirs (produced by the project). The survey is in the form of a small game and its data collected can be made available to the social workers and the project team.

d) Searching: a searching function for allowing the youths to quickly identify the information from (a), (b) or (e).

e) Dashboard: a sharing space for users to post/upload their digital products (games, videos, art works, etc.) for sharing. After attending training workshops available from the project, these products can be shared and users can rate them. After uploading, other users can access these products.

f) Quick Contact: this function supports a user to contact a social worker for assistance quickly. The user can enter free text or answer a number of pre-set questions and a social worker would be alerted and try to arrange help.

g) Instant Messaging: with different groups (among youths, parents, and social workers) formed, members within each group can do online chat and share information.

h) Check-In/Location Information: the function is to support users in contacting people nearby and using it as a social media platform for attracting youths. It allows a user to release his/her current location information to other users. It is a powerful tool so that a user can know any other users in the same area. Users can provide comments of the check-in locations. It is a voluntary basis and the access of the location information is limited to the people that the user allows. Some initial designs have been done as show in the next page.

The second application, Giai which focuses on parental support, is developed for parents with the functions below. For similar functions in Aurora, the functional descriptions are skipped.

a) Head News Centre

b) Searching

c) Dashboard

d) Quick Contact

e) Instant Messaging

f) Friend List: a function which allows a user to add a friend or create a new user group (limited to some types of users).

g) InfoSeek: it is a searching function similar to the Location Information function in the Youth Apps

h) Family Center: this is the intelligent alerting tool. It enables parents to input their observations of their children on a regular basis. The input is semi-guided which can help parents to enter structured information with certain set of unusual behaviors. Once parents' observations are entered, the second function is to perform analysis with the knowledge base of an expert system available in the backend server. The expert system is designed and developed by the project team with the help of anti-drug experts and social workers. In case of parent users do not how to start, a series of prompts (similar to a quiz/survey) can be used to collect data. The analysis result would be in form of hints or advice to parents, suggesting where to seek help or acquire further information. This function can work with (f) above to ease the worries of parents.

i) Parental Corner: it is a special sharing space for parental groups and social workers to share information in addition to Dashboard and Head News Center. Parents/social workers can post articles or provide comments/advice of any articles. This corner can access special cases, emergency procedures and FAQs that are stored in a knowledge repository developed by antidrug experts and/or social workers. A searching feature similar to (b) is available also. This function is aimed to encourage parents to form mutual supporting groups for helping each other.

j) Referral Center: the function is acting a help-contact service which allows a parent to have an initial contact with a social worker. This is a resourceful function for youths to receive answers of common questions, such as where and how they can seek help.

The third application, Theoi, is developed for social workers with the functions below. For similar functions in the other applications, the functional descriptions are skipped.

a) Head News Center

b) Searching

c) Dashboard

d) Instant Messaging

e) Friend List

f) InfoSeek

g) P-Action+: Parents are able to request help from social workers with the altering tool, an email with test records will be generated automatically and sent through their own email accounts which they used in registration.

h) Y-Action+: It is similar to P-Action+ but the corresponding contacts are youths.

i) Enrich+: According to the feedbacks from social workers, a web base platform is developed for this function, instead of included in the apps. It allows social workers to post and share information, and update the FAQ database.

Figure 7 shows the software architecture of the system in which different modules are labeled with different border colors. Modules in the red rectangles mean these are shared by all 3 applications. Modules in blue, green and purple rectangle represent the functions only serve students, parents and social workers respectively.

3.2. Behavior Analysis

For the behavior analysis component, there are 4 core modules as the training, classifying, calculating and collecting. In training module, we have chosen 2 techniques, decision tree (see Figure 9) and logistic regression, to build a classifier with the given training data. The calculating module is to estimate an index for possible explanation of the recommendation. The collecting module will receive user's feedback and store them for the enhancement of our analysis (i.e. re-training with additional data). The 4 modules will work together to analyze behavior of youths as shown in Figure 7.

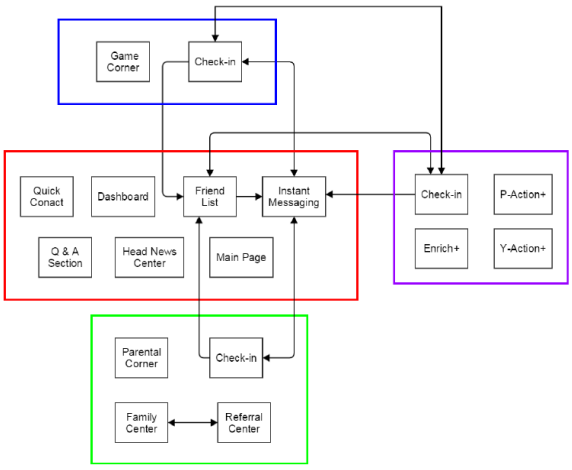


Figure 7. System modular design

The training module will run every time when the additional data is available. It will retrieve the training data and then apply a set of weight behavior criteria with the construction of a decision tree and the execution of logistic regression algorithm independently. The training results will pass to the classifying module and the calculating module afterwards. As the current available is small, the overhead of re-constructing the prediction model is acceptable. Yet, when the dataset size becomes larger, an incremental technique should be considered.

The classifying module is going to utilize the decision tree. A user is required to complete a set of guided questions which represent the behaviors of a potential drug abuser. Then the survey answers will be input to the decision tree. If it is classified as a non-drug abuser, it will bypass the calculating module and return the index as 0. Otherwise, it will go on the calculating module.

The calculating module is going to use the results of logistic regression. It is used for selecting a explanation description by estimating an index, which refers to the probability of drug abuse.

The determined class and index will return to the user. The collecting module is an optional module. A user can return expected status of the predicted drug abuse case to the server. This user feedback and the results of the previous survey will be added to the training data for later prediction improvement.

Input :
X = Training set,
A = each behavioral criteria scale for each training data set from database (training data)
Y = class for each training data set from database (training data)
B = each behaviora criterial scale from a user (according to survey)

Output: Decision tree

// Stage 1 – Training function (construct tree)
Create a new tree T with a single root node.
if stoppingCriterion(X) then
    Mark T as a leaf with the most common value in X as a label.
else
    Find a that obtain the best splittingCriterion(ai,X; y).
    Label n with a.
    for each outcome vi of a do
        Set subtree ti = treeGrowing(a=viX; A; y).
        Connect the root node of nt to the subtree ti with an edge that is labelled as vi.
    return T

Formula for spliting:

$$Q_{left}(\theta) = (x, y)|x_j <= t_m$$
$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta)$$
$$G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$$
$$\theta^* = \operatorname{argmin}_\theta G(Q, \theta)$$

// Stage 2 – Classifying function
    SearchTree(B, T)

Figure 8. Decision Algorithm

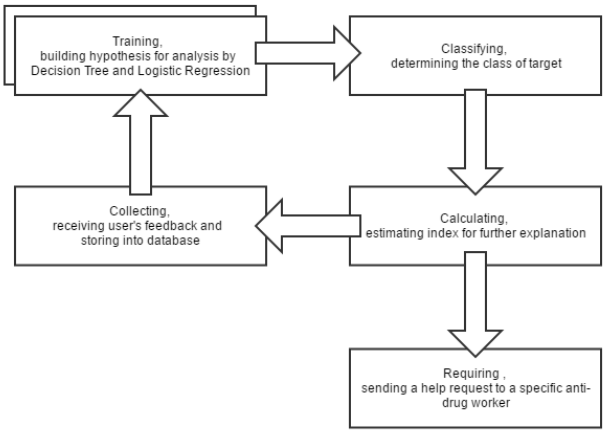


Figure 9. Behavior Analysis Framework

3.3. Altering

For the alerting component, there are 3 modules as confirmation, generation, and sending (as shown in Figure 10). In the confirmation module, we collect the information on the results of behavior analysis, user's demographic data, and the social worker whom users have chosen. Information would then be passed to the generation module. The generation module is used for generating email and instant message. They are generated by two servers in parallel. After that, the sending module will send the messages out.

3.1. Location Sharing and NearBy Support

Check-in support is a major component of our applications. Through the establishment of a healthy location-based social network (LBSN), it does not only share location-embedded information to an existing social network, but also create a new social structure by connecting individuals. With the location information, a person-to-person communication function can be enable, which is called the Friends Nearby Notification in our applications.

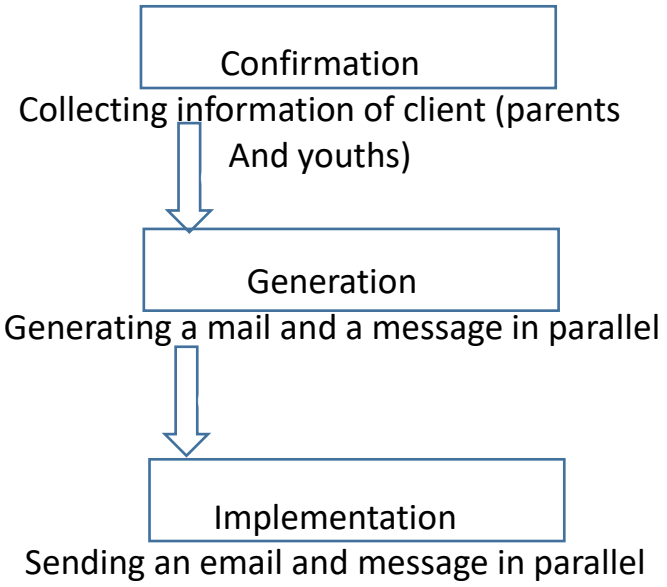


Figure 10. Alerting component

When a user would like to view others' comments to his current location, he can get them from a safe and secured database. Users can chat with their friends or other user after reading their comments. They can be redirected to Instant Messaging engine. In order to find their friends for having a face-to-face talk, the find route function is provided. The default origin will be the user's current location, and after user selecting a destination, data will be sent to the server. A string in JSON format will be returned and transformed into a polyline shown on the map which represents the route between two places. In our design, the Friends Nearby helps users to identify all nearby friends and accessibilities with the following features:

- Notify users when friends are in proximity
- Check nearby friends or accessibilities in regular time intervals
- Provide setting to turn off notification to support privacy
- Provide setting to hide from a particular friend.
- Show the nearby friends and accessibilities details, such as distance away from it, and contact info.

- Call or send short messages to nearby friends.

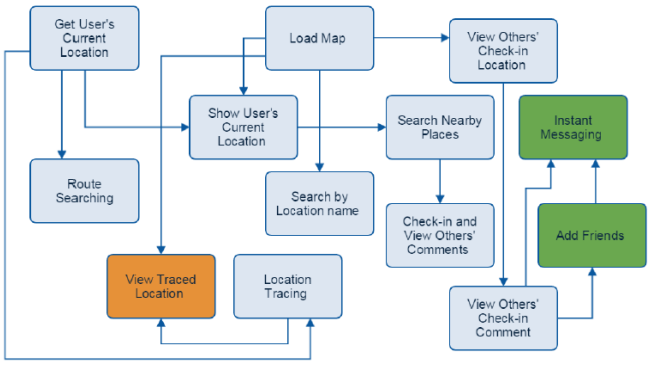


Figure 11. Check-in function

Figure 11 shows how the check-in function is designed and the starting module is the top left box in the figure, 'Get User Current Location.' The boxes in green are the functions that linked with other parts of the platform while the orange box is the function only available for social workers to trace locations of young people, provided with their permission. The locations of these young people will be uploaded to a central database for every 10 minutes. Social workers can view the students' location on the map by retrieving location information from the database.

4. Implementation

The mobile applications and its backend support are designed as a 3-tier client server model as shown in Figure 12. When a user uses a smartphone to access the system, requests will be sent to middle tier, i.e. web server, through HTTP request first. Then, the web server will translate the request into SQL queries and sent it to a database server to retrieve data. The return data will be transformed into JSON response and sent to the client device. All functions are based on different modules as shown in Figure 12.

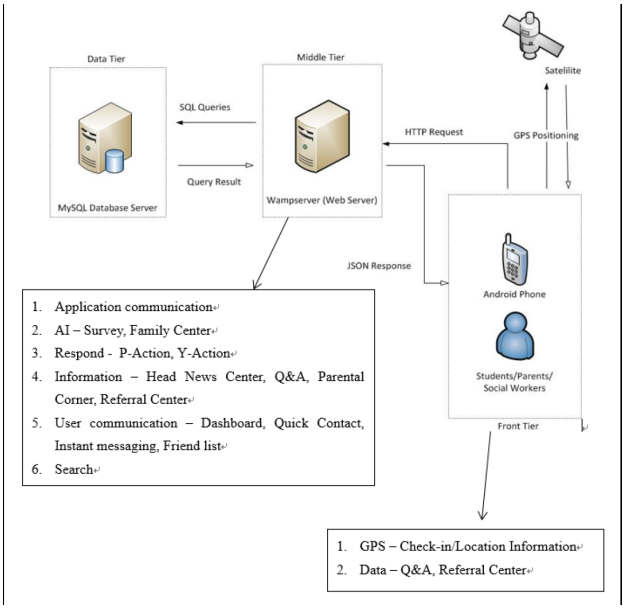


Figure12. Three tier implementation

In the current design, check-in requires only 2 relational tables. The first one stores basic user information, latitude and longitude, which are automatically uploaded by the application. The later one stores all check-in records. We do not archive all user location information since the amount of data is too large to be stored locally. For finding out a user's location, there are two common methods GPS and A-GPS. After balancing the advantages and disadvantages, we are using GPS to locate a user if it is available. If not, A-GPS will be applied. Figure 13 shows a location detection of a user.

Without GPS assistance, a user can check-in manually by entering his nearby location. Google Places API Web Service is used here since it can save device's resources. Several parameters including location, radius and API Access Key are required to retrieve the nearby locations. Location information in the format of latitude and longitude of user's current location can be gotten automatically by the system. Radius means the distance in meters within the returned place results. Check-in sets it as 100 meters since the known maximum error of A-GPS is 100 meters. After preparing these parameters properly, an HTTP request will be sent to Google. Maximum 20 locations will be returned by Google in JSON format. The returned results will contain sort of information about the location. Only place name will be shown to users in a list as shown in Figure 14.

Figure13. Location detection – Check in

Check-in provides location awareness support so that it can act as a channel to share, comment and rate any spots in different areas. Users can share their current location with others. Only 4 data values are required to "check-in", i.e. username, check-in's location, comment and rating. Username will automatically be gotten by the system. Check-in location is stored after user finds the nearby locations. A user can also enter comments and ratings about the location in Check-in page as shown in Figure 15. An error will be shown if the length of a comment is more than 160 English characters and 80 Chinese characters. A limit is set with the same idea from the famous social network like Twitter and Weibo. They have a special feature that the post length is limited. According to [17], 160 characters provide plenty of space to express most of their thoughts with friends. Chinese messages can usually present more information or feelings with fewer words.

Thus, 80 characters should be a proper length constrain in a check-in comment.

Figure14. A list of NearBy locations is returned

Figure15. Comments for a checked location

The behavior analysis module is implemented with Python 2.7 at the server end. We have selected the scikit-learn toolkit to support the decision tree and regression analysis [18]. The CART (Classification and Regression Trees) is used for building a model to determine if a person is a potential drug abuser. The same toolkit contains the logistic regression function in its 'liblinear' library and it is adopted.

The Drug Abuse Screening Test (DAST-10) is a 10-item brief screening tool formed the basis of the data being used in the behaviors modules [19]. Other data attributes, such as sleeping patterns, school academic performance, washroom visiting frequency and periods being alone, are included. During of the development of the module, the project team could not acquire the live data because of privacy issues. Instead, a pseudo dataset of 250 records (cases) is generated randomly with 20 attributes per case. Anti-drug experts were then invited to label the cases with different level of drug abuse (where 0 means nil and 5 means high potential). The first 200 records have been used for the decision tree and regression analysis in the scikit-learn toolkit. Cross-validation was then used to verify the accuracy and the results were in the range of 70-80%. The last 50 records were used as additional data for the continuous training. Yet, the results have not been improved.

The email alerting method is mainly based on PHPMailer. It is an open source class with methods to support HTML-based email and attachments. The messaging alerting method is mainly based on Messaging Queuing Telemetry Transport (MQTT) [20, 21]. MQTT is standardized by OASIS as standards to transport data. In MQTT, there is no message formats specialized in it and standard used is depending on developers. But it has advantages over using HTTP and XMPP, which it is low-weight in transmission as lower size in overhead and more energy saving [22, 23].

With the above set up environment, three mobile applications have been developed. The mobile apps for parents contain the alerting tool. It has a main page which displays Drug Abuse Prediction and or Result Feedback (see Figure 16).

In the behavior analysis, the parents act as observers to record their children's behaviors. When the parent starts the drug abuse prediction test, a series of questions on youths' behavior has to be answered in order to get the prediction results. The estimated probability of drug abuse is returned to the screen of the mobile device as shown in Figure 16. When a warning or alerting message is received, parents can seek help from social workers who have joined the platform according to their regions. In seeking the help, parents can provide more specific information about themselves and their children so that suitable social workers in the district can be located faster. They can also describe the situation in order to shorten the remedial time. When a help request is sent, the selected social workers will receive a notification from our platform as a reminder and email message describing parent's situation.

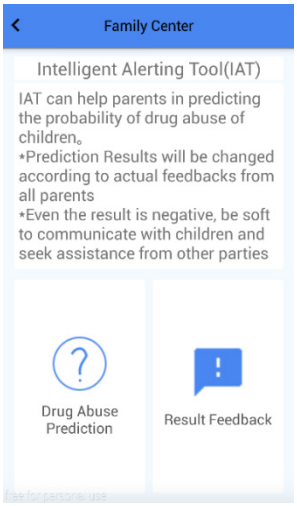


Figure 16a Main Page of behavior analysis

Figure 16b Result of probability of drug abuse

For result feedback, parents are required to answer "positive result" or "negative result" to confirm the drug abuse problem of their children (see Figure 17b). And the feedback is limited to previous detection, which any new prediction will overwrite the old predictions for feedback.

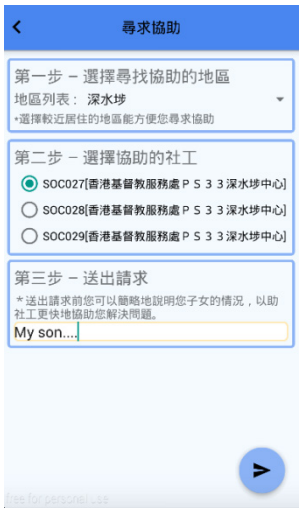


Figure 17a Seek help from social worker in different district

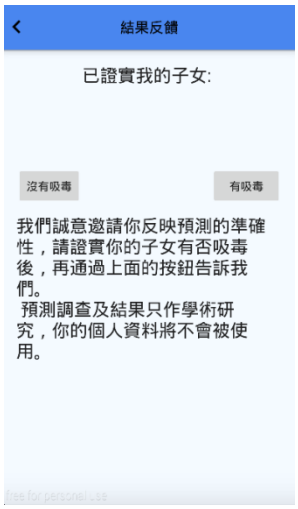


Figure 17b Result feedback

5. Conclusions

In this paper, we have presented three mobile applications with a set of intelligent tools for parents and social workers to help our young people as shown in Figure 18. The application for parents has a tool to analyze drug abuser behavior and send alert to them. The applications have been placed in GoogleStore for download availability.

The work is funded by the Beat Drugs Fund of HKSAR. We have invited a number of secondary schools, community centers, anti-drug experts and social workers to try the applications. All data is collected with a centralized server with confidentiality protection and agreement from users. Users can have options to select if any personal data can be used. Towards the end of the project, positive feedbacks from different user groups have been received. Many parents are interested in the alerting tool while expecting more rules and advice can be embedded.

Figure18. Three mobile applications for anti-drug abuse

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Isobe, 'On publication of denominational precious volumes (Jiao pai xi Pao juan) in Late Ming and Early Ch'ing as seen in, "Niwatazumi".', Studies of Publishing Culture in East Asia, vol. 8, pp. 187-226, 2004.

[2] H. Chan and T. Lo, 'Quality of Life of the Hidden Youth in Hong Kong', Applied Research Quality Life, vol. 9, no. 4, pp. 951-969, 2013.

[3] The Central Registry of Drug Abuse, "Central Registry of Drug Abuse Sixty-sixth Report," Narcotics Division, Hong Kong, 2016.

[4] Policy 21 Limited, "A review of estimation methods on prevalence of drug abuse Final Report," Narcotics Division, 2013.

[5] Social Surveys Section, "Thematic Household Survey Report No. 54 Information technology usage and penetration," Census and Statistics Department HKSAR, 2015.

[6] 禁毒資訊站 [Online]. Available: https://apkpure.com/%E7%A6%81%E6%AF%92%E8%B3%87%E8%A8%8A%E7%AB%99/com.antidrugs

[7] 許寶強, "校園驗毒：「認真」調查結果 馬虎禁毒措施," 明報, 2010.

[8] DrugOut [Online]. Available: https://itunes.apple.com/ao/app/dugout/id1163815036?mt=8

[9] Dihua Sun, Hong Luo, Liping Fu, Weining Liu, Xiaoyong Liao, and Min Zhao, "Predicting Bus Arrival Time on the Basis of Global Positioning System Data", Transportation Research Record: Journal of the Transportation Research Board,No. 2034, D.C., 2007, pp. 62–72.

[10] Depression Test [Online]. Available: https://www.psychologytoday.com/tests/health/depression-test

[11] " Cloud Messaging | Google Developers," Google Developers, 2015. [Online]. Available: https://developers.google.com/cloud-messaging/?hl=zh-tw.

[12] "Messenger for Android - Documentation - Facebook for Developers," Facebook Developers, [Online]. Available: https://developers.facebook.com/docs/messenger/android.

[13] Hani Badran, Pierre Pluye, Roland Grad, "Advantages and Disadvantages of Educational Email Alerts for Family Physicians", J Med Internet Res 2015;17(2):e49.

[14] BradleyTony. (2012). Email vs. IM vs. SMS: Choosing the Right One. PCWorld.

[15] Google Maps will let you share your location with friends and family for a specific period of time [Online]. Available:,https://techcrunch.com/2017/03/22/google-maps-now-lets-you-share-your-location-with-friends-and-family-for-a-specific-period-of-time/

[16] "Google Map," Google, [Online]. Available: https://www.google.com.hk/maps?source=tldsi&hl=zh-TW.

[17] "Find My Friends," Apple Inc., [Online]. Available: https://www.apple.com/apps/find-my-friends/.

[18] Scikit learn [Online]. Available: http://scikit-learn.org/stable/

[19] The Drug Abuse Screening Test (DAST-10) [Online]. Available: https://www.drugabuse.gov/sites/default/files/dast-10.pdf

[20] "MQTT," Mqtt.org, 2015. [Online]. Available: http://mqtt.org.

[21] Y. Zhu, "Mobile Application Message Push and MQTT Protocol," in Wireless Internet Technology, vol. 8, 2015, pp. 1-3.

[22] "The XMPP Standards Foundation," Xmpp.org, 2015. [Online]. Available: http://xmpp.org.

[23] W. Huang, "Design of Android Instant Messaging system based on XMPP agreement," in International Electronic Elements , vol. 19, no. 8, 2011, pp. 57-59.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 443-450 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Actuator Fault Reconstruction based Adaptive Polytopic Observer for a Class of Continuous-Time LPV Systems

Radhia Houimli[*,1,2], Neila Bedioui[1], Mongi Besbes[2]

[1]*Robotics, Informatics and Complex Systems (RISC), ENIT, University of Tunis Al Manar, Institute, 1002, Tunisia*

[2]*Higher Institute of Information and Communication Technologies, University of Carthage, 1164, Tunisia.*

A R T I C L E I N F O

Article history:
Received: 04 December, 2017
Accepted: 07 January, 2018
Online: 10 February, 2018

Keywords :
Polytopic Linear Parameter-Varying (LPV) system
Adaptive observer
LMI

A B S T R A C T

This paper is an extension of work originally presented in conference name. The goal is to propose new fault detection and fault isolation techniques for a polytypic linear parameter-varying system (LPV). In this work, an adaptive observer design is formulated for a given polyquadratic Lyapunov function. Subsequently, new sufficient conditions are given in terms of Linear Matrix Inequalities (LMIs).
To show the effectiveness of the proposed algorithm, an illustrative example is included.

1. Introduction

Research on fault detection, from a theoretical and experimental point of view, has been intensively developed during the last decades. All physical systems must operate normally and without anomaly. However, some conditions cause one or more faults in the process and interrupt this operation. So fault detection is an essential task to avoid degradation of system performance or even its damage [1].

The synthesis of observers for uncertain systems is based on the asymptotic stability of the error estimation equations or on the eliminating influence of uncertain perturbations and measurement of error on errors estimation. Linear Parameter varying system (also known as LPV system) is a special class of system that includes parameters-varying equations and parameters-varying state-space equations. These uncertain systems can be considered as a linearization of state-space nonlinear systems.

The polytopic LPV form is a special class of LPV systems. Indeed, it is a description of the system as a convex combination of sub-models defined by the vertices of a convex polytope [2] [3]. Subsequently, these sub-models are combined by convex weighting functions which give a global model. Similarly, [4] proposes a linear piecewise interpolation model of a diesel engine. The nonlinear model of the machine was transformed into an LPV model.

The synthesis of LPV observers is a direct extension of the LTI control methodologies [5]. The LPV theory has allowed to extend linear methods to nonlinear domains [6]. LPV modeling is used to study nonlinear systems, multiple models or switched models [7].

In this paper, the main contribution is a generalization of the obtained results for actuator fault reconstruction in [1] for LTI systems to continuous-time LPV systems. It was proved that linear methods could be extended to nonlinear domains [8] [9]. In this direction, the actuator fault reconstruction problem is articulated as an LMI feasibility problem. The existence of polyquadratic Lyapunov function could insure the stability of the error estimation [10].

Due to the varying parameter, the polyquadratic approach considers that the Lyapunov function depends on the parameters associated with the description of the polytope.

The paper is organized as follows: In section 2 a model of linear polytopic time-varying (LPV) system is presented. An adaptive observer for fault detection is described in section 3. In section 4, we introduce the polyquadratic adaptive observer for the polytopic LPV system which leads to less conservative conditions on terms of LMI. The simulation result illustrates the effectiveness of our contribution.

Notation. For conciseness the following notations are used: $sym\ (A) = A + A^T$, $\begin{bmatrix} A & B \\ \bullet & C \end{bmatrix} =, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$

[*]Radhia Houimli, BP 8, Face Street, 2021, Tunisia. radhia.houimli@gmail.com

2. Problem Statements and Preliminaries

A continuous-time LPV system in presence of fault can be described by the state-space equations in the following form:

$$\begin{cases} \dot{x}(t) = A(\theta(t))x(t) + B(\theta(t))u(t) + E(\theta(t))f(t) \\ y(t) = Cx(t) \end{cases} \quad (1)$$

Where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$ and $y \in \mathbb{R}^p$ are, respectively, the state space, the input and the output vectors of the system. Variable $f(t)$ represents actuator fault.

The scheduling θ is a set of varying parameters evaluated in hypercube domain Θ such as:

$$\Theta = \left\{ \theta(t) \in \mathbb{R}^p \mid \theta_1 \in \left[\theta_1^{\min}, \theta_1^{\max}\right], ..., \theta_p \in \left[\theta_p^{\min}, \theta_p^{\max}\right] \right\} \quad (2)$$

Where $\theta_i^{\min}$ and $\theta_i^{\max}$, $i = 1,...,p$ are the lower and upper bounds of the parameter.

The parameter dependent Lyapunov function is assumed to be measurable [11] :

$$\theta(t) = \left[\theta_1(t) \quad \theta_2(t) \quad \cdots \quad \theta_r(t)\right]^T \in \mathbb{R}^r \quad (3)$$

Furthermore, LPV system (1) can be defined via barycentric combination of a matrix polytope described by N vertices [12].

$$\begin{aligned} A(\theta(t)) &= \sum_{i=0}^{N} \rho_i(\theta(t)) A_i \\ B(\theta(t)) &= \sum_{i=0}^{N} \rho_i(\theta(t)) B_i, \\ E(\theta(t)) &= \sum_{i=0}^{N} \rho_i(\theta(t)) E_i \end{aligned} \quad (4a)$$

$$\begin{aligned} &\sum_{i=0}^{N} \rho_i(\theta(t)) = 1 \\ &\rho_i \geq 0 \end{aligned} \quad (4b)$$

Where $A_i \in \mathbb{R}^{n \times n}$, $B_i \in \mathbb{R}^{n \times m}$, $E_i \in \mathbb{R}^{n \times r}$ are time invariant matrices defined for the i^{th} ertex of the hypercube and $\rho_i\left((\theta(t))\right) = \rho(\theta_i^{min}, \theta_i^{max}, \theta, t)$. The weighting function $\rho_i\left((\theta(t))\right)$ define the relative contribution of each vertices (A_i, B_i, E_i) to build the system described by $(A(\theta), B(\theta), E(\theta))$.

2.1. Assumption 1 [11]:

The state-space matrices $\left(A\left(\rho(\theta(t))\right), B\left(\rho(\theta(t))\right)\right)$ are continuous and bounded functions and depend on $\rho(\theta(t))$.

2.2. Assumption 2 [11]:

The real parameters $\rho(\theta(t))$ that can be known by on-line measurement values exist in LPV system and vary in a polytope Θ as:

$$\begin{aligned} &\rho(t) \in \Theta \\ &\Theta = \left\{ \sum_{i=1}^{N} \alpha_i(t) w_i : \alpha_i(t) \geq 0, \sum_{i=1}^{N} \alpha_i(t) = 1, N = 2^r \right\} \end{aligned} \quad (5)$$

And the rate of variation $\dot{\rho}(\theta(t))$ are well defined at all times and vary in a polytope $\Theta_v$as:

$$\dot{\rho}(t) \in \Theta_v \quad (6)$$

With

$$\Theta_v = \left\{ \sum_{k=1}^{N} \beta_k(t) v_k : \beta_k(t) \geq 0, \sum_{k=1}^{N} \beta_k(t) = 1, N = 2^r \right\} \quad (7)$$

The following assumption is made:

2.3. Assumption 3:

Without loss of generality matrix C is considered full row rank.

2.4. Assumption 4:

$$rank(CE_i) = rank(E_i) = p \qquad \forall i = [1,...,N] \quad (8)$$

2.5. Assumption 5:

The triple matrix (A_i, E_i, C) is observable.

For simplicity, the time variable t of $\theta(t)$ will be omitted if no confusion is caused.

3. Adaptive Observer Design

For polytopic LPV system (1), an adaptive observer is described by the following state representation:

$$\begin{cases} \dot{\hat{x}}(t) = A(\theta(t))\hat{x}(t) + B(\theta(t))u(t) \\ \qquad + E(\theta(t))\hat{f}(t) - L(\theta(t))(y(t) - \hat{y}(t)) \\ \hat{y}(t) = C\hat{x}(t) \end{cases} \quad (9)$$

Where $\hat{x}(t)$, $\hat{y}(t)$ are the state and outputs estimated vectors and $\hat{f}(t)$is the fault estimation.

In this case the gain matrix is given by the following polytopic form:

$$L(\theta) = \sum_{i=0}^{N} \rho_i(\theta) L_i \quad (10)$$

Where L_i represents the gain of i^{th}vertex.

Remark 1: Since it has been assumed that the pair (A_i, C) is observable, the gain matrices L_i can be selected such that $(A_i - L_i C)$ is stable.

Denote $e_x(t)$, $e_y(t)$, $e_f(t)$ are respectively state ,output and fault estimations errors :

$$\begin{aligned} e_x(t) &= \hat{x}(t) - x(t) \\ e_y(t) &= \hat{y}(t) - y(t) \\ e_f(t) &= \hat{f}(t) - f(t) \end{aligned} \quad (11)$$

Then, the error dynamics are expressed as follows:

$$\dot{e}_x(t) = \sum_{i=1}^{N} \rho_i(\theta(t))\left((A_i - L_i C)e_x(t) + E_i e_f(t)\right) \quad (12)$$

$$\dot{e}_y(t) = C\dot{e}_x(t) \quad (13)$$

The default $f(t)$ is constant, hence $\dot{f}(t) = 0$ [13], consequently the derivate of $e_f(t)$ with respect to time can be written as:

$$\dot{e}_f(t) = \dot{\hat{f}}(t) \tag{14}$$

The state observer (9) is combined with the law for the fault estimation updating of the form [1]

$$\dot{\hat{f}}(t) = -\Gamma F e_y(t) \tag{15}$$

Where $F \in \mathbb{R}^{r\times p}$ and $\Gamma \in \mathbb{R}^{r\times r}$ is the learning rate.

It has been to note a modification of (13) is presented in [1], [6] for time varying $f(t)$ in the form:

$$\dot{\hat{f}}(t) = -\Gamma F(\theta)\left(\dot{e}_y(t) + \sigma e_y(t)\right) \tag{14}$$

Where $\sigma \in \mathbb{R}$ is a positive scalar and can guaranty $\lim_{t\to\infty} e_x(t) = 0$ and $\lim_{t\to\infty} e_f(t) = 0$.

4. Main Result

Consider the LPV system described by (1) with an additive fault. The choice of the parameter-dependent Lyapunov functions for polytopic systems is a dilemma in the literature. The rate of change of the scheduling parameter could be represented in diverse methods. [14] proposed a rate of change that cannot be physically justified. Furthermore, in the expression given by [13], the derivative of the uncertain parameter does not impose special conditions.

In this section, we develop a new adaptive observer for LPV polytopic system. Before, we introduce some instrumental tools which will be used in the proof of characterization of this observer.

4.1. Lemma 1 [1]:

Given scalar $\mu > 0$ and symmetric positive definite matrix $P\left(\rho(\theta(t))\right)$, the following inequality holds:

$$2x^T y \le \frac{1}{\mu} x^T P(\theta) x + \mu y^T P(\theta)^{-1} y \qquad x, y \in \mathbb{R}^n \tag{17}$$

4.2. Lemma 2 *[15]:*

Given a symmetric matrix $\psi \in \mathbb{R}^{n\times n}$, and two matrices P, Q of column dimensions n, there exists X such that the following LMI holds:

$$\psi + sym(P^T X^T Q) < 0 \tag{18}$$

If and only if the projection inequalities with respect to X are satisfied:

$$\mathcal{N}_P \psi \mathcal{N}_P^T < 0, \quad \mathcal{N}_Q^T \psi \mathcal{N}_Q < 0 \tag{19}$$

Where $\mathcal{N}_p$ and $\mathcal{N}_Q$ denote arbitrary bases of the null spaces of P and Q respectively.

Proof. See [15]. ■

4.3. Lemma 3:

Let Φ a symmetric matrix and N, J matrices of appropriate dimensions. The following statements are equivalent:

- $\Phi < 0$ and $\Phi + NJ^T + JN^T < 0$.
- There exists a matrix X such that:

$$\begin{bmatrix} \Phi & J + NX \\ J^T + X^T N^T & -X - X^T \end{bmatrix} < 0 \tag{20}$$

Proof: The proof is obtained remarking that (19) can be developed as follows:

$$\begin{bmatrix} \Phi & J + NX \\ J^T + X^T N^T & -X - X^T \end{bmatrix} = \begin{bmatrix} \Phi & J \\ J^T & 0 \end{bmatrix} + sym\left\{ \begin{bmatrix} 0 \\ I \end{bmatrix} X^T \begin{bmatrix} N^T & -I \end{bmatrix} \right\} < 0 \tag{21}$$

and by applying Lemma 2 . ■

In this part, we consider the case such as the fault is time-varying, which implies $\dot{f}(t) \neq 0$, and the derivate of $e_f(t)$ with respect to time is:

$$\dot{e}_f(t) = \dot{\hat{f}}(t) - \dot{f}(t) \tag{22}$$

The objective of this section is to propose an approach to design a new adaptive observer for polytopic LPV system (1). So, we propose the following Theorem 2.

Theorem 2. Under the assumptions 1, 2 and 3, the system (9) is an adaptive observer for the system (1) if, for a given scalars $\sigma > 0$, $\mu > 0$, $\alpha > 0$ and $\nu > 0$, there exists, for each vertex, asymmetric positive definite matrix $P_i \in \mathbb{R}^{n\times n}$, $L_i \in \mathbb{R}^{n\times r}$, $G_i \in \mathbb{R}^{r\times r}$ and $X_1 \in \mathbb{R}^{n\times n}$ such that the following conditions hold:

$$\begin{bmatrix} \nu P_i - 2P_i + sym(P_i A_j) & -\frac{2}{\sigma} A_j^T P_i E_k & P_i + X_1 & P_i - \alpha C^T L_i^T \\ \bullet & -\frac{2}{\sigma} E_k^T P_i E_j + \frac{1}{2\mu\sigma} G_i & 0 & -\frac{2}{\sigma} E_j^T P_i \\ \bullet & \bullet & -X_1 - X_1^T & 0 \\ \bullet & \bullet & \bullet & -2\alpha I \end{bmatrix} < 0 \tag{23}$$

Under constraint:

$$E_j^T P_i = F_i C \tag{24}$$

The gain observer matrix of the system (1) is given by the following polytopic form:

$$L(\theta) = \sum_{i=0}^{N} \rho_i(\theta) L_i \tag{25}$$

Remark 3: The principle of the polytopic formulation is based on the fact that the system and stability conditions (here in a LMI form) have affine dependence on the parameters. If, for some reason, the affine dependence is lost, the stability of the system is not equivalent (or even implied only) to the feasibility of the LMI at each vertex [16], [17], [18].

Proof: With respecting to the system parameter, it is clear that $e_x(t)$ is linear. Thereby, consider the Lyapunov polytopic function defined by:

$$V\left(e_x(t),\right) e_f(t)) = e_x^T(t) P(\theta) e_x(t) + e_f^T(t) \Gamma^{-1} e_f(t) \tag{26}$$

Where $P\left(\rho(\theta(t))\right) > 0$ is a symmetric positive defined matrix.

Then, the derivative of (26) with respect to t is:

$$\dot{V}\left(e_x(t), e_f(t)\right) = \dot{V}_1\left(e_x(t), e_f(t)\right) + \dot{V}_2\left(e_x(t), e_f(t)\right) < 0 \quad (27)$$

Where :

$$\begin{aligned}\dot{V}_1\left(e_x(t), e_f(t)\right) &= \dot{e}_x^T(t)P(\theta)e_x(t) \\ &+ e_x^T(t)\dot{P}(\theta)e_x(t) + e_x^T(t)P(\theta)\dot{e}_x(t) \\ &= e_x^T(t)\Big[\dot{P}(\theta) + \left(A(\theta) - L(\theta)C\right)^T P(\theta) \\ &+ P(\theta)\left(A(\theta) - L(\theta)C\right)\Big]e_x(t) \\ &+ e_f^T(t)E^T(\theta)P(\theta)e_x(t) \\ &+ e_x^T(t)P(\theta)E(\theta)e_f(t)\end{aligned} \quad (28)$$

And

$$\begin{aligned}\dot{V}_2\left(e_x(t), e_f(t)\right) &= \frac{1}{\sigma}\dot{e}_f^T(t)\Gamma^{-1}e_f(t) + \frac{1}{\sigma}e_f^T(t)\Gamma^{-1}\dot{e}_f(t) \\ &= \frac{1}{\sigma}\left(\dot{\hat{f}}(t) - \dot{f}(t)\right)^T \Gamma^{-1}e_f(t) \\ &+ \frac{1}{\sigma}e_f^T(t)\Gamma^{-1}\left(\dot{\hat{f}}(t) - \dot{f}(t)\right)\end{aligned} \quad (29)$$

Substituting (16) into (28) leads to :

$$\begin{aligned}\dot{V}_2\left(e_x(t), e_f(t)\right) &= -\frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\dot{f}(t) \\ &- \frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\Gamma\Big[F(\theta)\left(\dot{e}_y(t) + \sigma e_y(t)\right)\Big] \\ &= -\frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\dot{f}(t) - \frac{2}{\sigma}e_f^T(t)F(\theta)C\dot{e}_x(t) \\ &- e_f^T(t)F(\theta)Ce_x(t) - e_x^T(t)C^T F^T(\theta)e_f(t)\end{aligned} \quad (30)$$

Then, substituting (12) and (13) into (30), the following inequality is hold:

$$\begin{aligned}\dot{V}\left(e_x(t), e_f(t)\right) &= e_x^T(t)\Big[\dot{P}(\theta) + \left(A(\theta) - L(\theta)C\right)^T P(\theta) \\ &+ P(\theta)\left(A(\theta) - L(\theta)C\right)\Big]e_x(t) \\ &+ e_f^T(t)E^T(\theta)P(\theta)e_x(t) \\ &+ e_x^T(t)P(\theta)E(\theta)e_f(t) \\ &- e_f^T(t)F(\theta)Ce_x(t) \\ &- e_x^T(t)C^T F^T(\theta)e_f(t) \\ &- \frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\dot{f}(t) - \frac{2}{\sigma}e_f^T(t)F(\theta)C\dot{e}_x(t)\end{aligned} \quad (31)$$

If the following condition is introduced:

$$\begin{aligned}&e_f^T(t)\Big[E^T(\theta)P(\theta) - F(\theta)C\Big]e_x(t) \\ &+ e_x^T(t)\Big[P(\theta)E(\theta) - C^T F^T(\theta)\Big]e_f(t) = 0\end{aligned} \quad (32)$$

This implies that:

$$E^T(\theta)P(\theta) = F(\theta)C \quad (33)$$

The inequality (31) becomes by using (33) and :

$$\begin{aligned}\dot{V}\left(e_x(t), e_f(t)\right) &= e_x^T(t)\Big[\dot{P}(\theta) + \left(A(\theta) - L(\theta)C\right)^T P(\theta) \\ &+ P(\theta)\left(A(\theta) - L(\theta)C\right)\Big]e_x(t) \\ &- \frac{2}{\sigma}e_f^T(t)E^T(\theta)P(\theta)\left(A(\theta) - L(\theta)C\right)e_x(t) \\ &- \frac{2}{\sigma}e_f^T(t)E^T(\theta)P(\theta)E(\theta)e_f(t) \\ &- \frac{2}{\sigma}e_f^T(t)\Gamma^{-1}\dot{f}(t)\end{aligned} \quad (34)$$

From Lemme 1, we can suppose that:

$$\begin{aligned}&2\left(-\frac{1}{2\sigma}e_f(t)\right)^T\left(\Gamma^{-1}\dot{f}(t)\right) \\ &\le \frac{1}{2\mu\sigma}e_f^T(t)Ge_f(t) + \frac{\mu}{2\sigma}f_1^2\lambda_{\max}\left(\Gamma^{-1}G^{-1}\Gamma^{-1}\right)\end{aligned} \quad (35)$$

Then, subsisting (35) in (34), the following inequality is done:

$$\begin{bmatrix} a_{11} & * \\ a_{21} & a_{22}\end{bmatrix} < 0 \quad (36)$$

With:

$$a_{11} = \dot{P}(\theta) + sym\left(\left(A(\theta) - L(\theta)C\right)^T P(\rho)\right)$$

$$a_{21} = -\frac{2}{\sigma}E^T(\theta)P(\theta)\left(A(\theta) - L(\theta)C\right)$$

$$a_{22} = -\frac{2}{\sigma}E^T(\theta)P(\theta)E(\theta) + \frac{1}{2\mu\sigma}G(\theta)$$

The derivate of the Lyapunov function is defined as follows:

$$dP(\theta)/dt = \sum_{k=1}^{N}\beta_k(t)Pt(v_k) = \sum_{k=1}^{N}\beta_k(t)\left(P(v_k) - \hat{P}_0\right) \quad (37)$$

$$\dot{P}(\theta) = \sum_{i=1}^{N}\dot{\rho}_i(\theta)P_i \quad (38)$$

$$\sum_{i=1}^{N}\dot{\rho}_i = 0 \quad (39)$$

The rate $\dot{\rho}(t)$ can be represented in several ways. In fact most of the time, it is difficult to give adequate modeling of it. For LPV system, the derived parameter does not vanish as in the LTI case.

In our case, we suppose that [14]:

$$\dot{\rho}(t) < \upsilon\rho(t) \quad (40)$$

$$\dot{P}(\theta) < \upsilon P(\theta) \quad (41)$$

Unfortunately, (41) is not convex in P and L, and cannot be solved by the LMI tools.

We can introduce some transformations to simplify the product term $(P(\theta)L(\theta)C)$ of the inequality (41). In fact, in this solution

we introduce an additive variable in order to allow the decoupling between the Lyapunov matrix and the observer gain in one side and to preserve a general structure to the Lyapunov matrix in the other side.

We suppose that:

$$\Phi=\begin{bmatrix} \upsilon P(\theta)-2P(\theta)+sym\left(P(\theta)A(\theta)\right) & \\ -\frac{2}{\sigma}E^T(\theta)P(\theta)A(\theta) & \\ * & \\ -\frac{2}{\sigma}E^T(\theta)P(\theta)E(\theta)+\frac{1}{2\mu\sigma}G(\theta) \end{bmatrix} \quad (42)$$

$$N^T=\begin{bmatrix} I & 0 \\ -L(\theta)C & 0 \end{bmatrix} \quad (43)$$

$$J=\begin{bmatrix} P(\theta) & P(\theta) \\ 0 & -\frac{2}{\sigma}E^T(\theta)P(\theta) \end{bmatrix} \quad (44)$$

By lemma 2 with (42), (43) and (44), there exists a matrix X of appropriate dimensions such that inequality (45) is satisfied.

$$\begin{bmatrix} M_1 & M_2 & M_3 & M_4 \\ * & M_5 & 0 & M_6 \\ * & * & -X_1-X_1^T & 0 \\ * & * & 0 & -2\alpha I \end{bmatrix}<0 \quad (45)$$

Where

$$\mathrm{X}=\begin{bmatrix} X_1 & 0 \\ 0 & \alpha I \end{bmatrix} \quad (46)$$

$$M_1=\upsilon P(\theta)-2P(\theta)+sym(P(\theta)A(\theta))$$

$$M_2=-\frac{2}{\sigma}A(\theta)P(\theta)E(\theta)$$

$$M_3=P(\theta)+X_1$$

$$M_4=P(\theta)-\alpha C^T L^T(\theta)$$

$$M_5=-\frac{2}{\sigma}E^T(\theta)P(\theta)E(\theta)+\frac{1}{2\mu\sigma}G(\theta)$$

$$M_6=-\frac{2}{\sigma}E^T(\theta)P(\theta)$$

Remark 4: The main advantage of problem (45) will appear when dealing with poly-quadratic observer. In that case, we will see that it theoretically improves the obtained results.

5. Numerical example

The above-described algorithm was applied to reconstruct the fault applied to the following LPV system described in [5] as:

$$A(\theta)=\begin{bmatrix} -1.75+\theta_2 & 1 & 0 & 0 \\ 1 & -1+\theta_1 & 0 & 0 \\ -1.8 & -1 & -0.75+\theta_1 & 0 \\ -1 & 0 & 0 & -1-\theta_2 \end{bmatrix}$$

$$B(\theta)=\begin{bmatrix} 1+\theta_1 & 1 \\ 1 & 0.5+\theta_2 \\ 1 & 0 \\ \theta_2 & 0 \end{bmatrix},\ C=\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$E(\theta)=\begin{bmatrix} 0 \\ 0.6+\theta_1 \\ 0 \\ 1 \end{bmatrix}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\rho_i(\theta)\rho_j(\theta)\rho_k(\theta)\begin{bmatrix} vP_i-2P_i+sym(P_iA_k) & \cdots & -\frac{2}{\sigma}A_jP_iE_k & \cdots \\ * & \cdots & -\frac{2}{\sigma}E_k^TP_iE_j+\frac{1}{2\mu\sigma}G_i & \cdots \\ * & \cdots & * & \cdots \\ * & \cdots & * & \cdots \end{bmatrix}$$

$$\begin{bmatrix} P_i+X_1 & P_i-\alpha C^TL_i^T \\ 0 & -\frac{2}{\sigma}E_j^TP_i \\ -X_1-X_1^T & 0 \\ 0 & -2\alpha I \end{bmatrix}<0 \quad (47)$$

The gain scheduling vector is defined as:

$$\theta=\begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}^T \quad (48)$$

Where

$$\theta_1\in\begin{bmatrix} -0.05 & 0.05 \end{bmatrix} \quad (49)$$

$$\theta_2\in\begin{bmatrix} -0.1 & 0.1 \end{bmatrix} \quad (50)$$

The system has four vertices and evolves in a hypercube. The weighting functions, which verified (4b), are computed as follows:

$$\begin{aligned} \rho_1(\theta)&=\frac{\theta_1-\underline{\theta_1}}{\overline{\theta_1}-\underline{\theta_1}}\frac{\theta_2-\underline{\theta_2}}{\overline{\theta_2}-\underline{\theta_2}}=\frac{(\theta_1+0.05)(\theta_2+0.1)}{0.02} \\ \rho_2(\theta)&=\frac{\theta_1-\underline{\theta_1}}{\overline{\theta_1}-\underline{\theta_1}}\frac{\overline{\theta_2}-\theta_2}{\overline{\theta_2}-\underline{\theta_2}}=\frac{(\theta_1+0.05)(0.1-\theta_2)}{0.02} \\ \rho_3(\theta)&=\frac{\overline{\theta_1}-\theta_1}{\overline{\theta_1}-\underline{\theta_1}}\frac{\theta_2-\underline{\theta_2}}{\overline{\theta_2}-\underline{\theta_2}}=\frac{(0.05-\theta_1)(\theta_2+0.1)}{0.02} \\ \rho_4(\theta)&=\frac{\overline{\theta_1}-\theta_1}{\overline{\theta_1}-\underline{\theta_1}}\frac{\overline{\theta_2}-\theta_2}{\overline{\theta_2}-\underline{\theta_2}}=\frac{(0.05-\theta_1)(0.1-\theta_2)}{0.2} \end{aligned} \quad (51)$$

The four local models represented the LPV system are calculated as the following:

$$A_1=\begin{bmatrix} -1.85 & 1 & 0 & 0 \\ -1 & -1.05 & 0 & 0 \\ -1.8 & -1 & -0.8 & 0 \\ -1 & 0 & 0 & -1.1 \end{bmatrix} \quad (52)$$

$$A_2 = \begin{bmatrix} -1.65 & 1 & 0 & 0 \\ -1 & -1.05 & 0 & 0 \\ -1.8 & -1 & -0.8 & 0 \\ -1 & 0 & 0 & -0.9 \end{bmatrix} \quad (53)$$

$$A_3 = \begin{bmatrix} -1.85 & 1 & 0 & 0 \\ -1 & -0.95 & 0 & 0 \\ -1.8 & -1 & -0.7 & 0 \\ -1 & 0 & 0 & -1.1 \end{bmatrix} \quad (54)$$

$$A_4 = \begin{bmatrix} -1.65 & 1 & 0 & 0 \\ -1 & -0.95 & 0 & 0 \\ -1.8 & -1 & -0.7 & 0 \\ -1 & 0 & 0 & -0.9 \end{bmatrix} \quad (55)$$

$$B_1 = \begin{bmatrix} 0.95 & 1 \\ 1 & 0.4 \\ 1 & 0 \\ -0.1 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0.95 & 1 \\ 1 & 0.6 \\ 1 & 0 \\ 0.1 & 0 \end{bmatrix}, \quad (56)$$

$$B_3 = \begin{bmatrix} 1.05 & 1 \\ 1 & 0.4 \\ 1 & 0 \\ -0.1 & 0 \end{bmatrix}, \quad B_4 = \begin{bmatrix} 1.05 & 1 \\ 1 & 0.6 \\ 1 & 0 \\ 0.1 & 0 \end{bmatrix} \quad (57)$$

$$E_1 = E_2 = \begin{bmatrix} 0 \\ 0.55 \\ 0 \\ 1 \end{bmatrix}, \quad E_3 = E_4 = \begin{bmatrix} 0 \\ 0.65 \\ 0 \\ 1 \end{bmatrix} \quad (58)$$

By applying algorithm (9), gains matrices are as the following:

$$L_1 = 10^{-7} \times \begin{bmatrix} 0.8103 & 0.1578 & -0.489 \\ 0.1433 & 0.0788 & -0.0832 \\ 0.3573 & 0.0560 & -0.1950 \\ -0.3956 & 0.0074 & 0.2464 \end{bmatrix}$$

$$L_2 = 10^{-7} \times \begin{bmatrix} 0.9953 & 0.2005 & -0.6049 \\ 0.1606 & 0.0981 & -0.0962 \\ 0.4209 & 0.0641 & -0.2329 \\ -0.4902 & 0.0192 & 0.3089 \end{bmatrix}$$

$$L_3 = 10^{-7} \times \begin{bmatrix} 0.7851 & 0.1489 & -0.4744 \\ 0.1376 & 0.0795 & -0.0798 \\ 0.3467 & 0.0535 & -0.1886 \\ -0.3845 & 0.0104 & 0.2398 \end{bmatrix}$$

$$L_4 = 10^{-7} \times \begin{bmatrix} 0.8443 & 0.1676 & -0.5104 \\ 0.1469 & 0.0780 & -0.0860 \\ 0.3687 & 0.0576 & -0.2022 \\ -0.4115 & 0.0053 & 0.2563 \end{bmatrix}$$

$$P_1 = 10^{-7} \times \begin{bmatrix} 0.1529 & 0.0034 & 0.0439 & -0.0329 \\ 0.0034 & 0.0098 & 0.0042 & -0.0004 \\ 0.0439 & 0.0042 & 0.0175 & -0.0138 \\ -0.0329 & -0.0004 & -0.0138 & 0.0254 \end{bmatrix}$$

$$P_2 = 10^{-7} \times \begin{bmatrix} 0.1877 & 0.0023 & 0.0523 & -0.0407 \\ 0.0023 & 0.0126 & 0.0044 & 0.0002 \\ 0.0523 & 0.0044 & 0.0203 & -0.0169 \\ -0.0407 & 0.0002 & -0.0169 & 0.0328 \end{bmatrix}$$

$$P_3 = 10^{-7} \times \begin{bmatrix} 0.1500 & 0.0032 & 0.0429 & -0.0319 \\ 0.0032 & 0.0097 & 0.0040 & -0.0000 \\ 0.0429 & 0.0040 & 0.0172 & -0.0134 \\ -0.0319 & -0.0000 & -0.0134 & 0.0250 \end{bmatrix}$$

$$P_4 = 10^{-7} \times \begin{bmatrix} 0.1562 & 0.0033 & 0.0448 & -0.0343 \\ 0.0033 & 0.0099 & 0.0043 & -0.0008 \\ 0.0448 & 0.0043 & 0.0179 & -0.0144 \\ -0.0343 & -0.0008 & -0.0144 & 0.0261 \end{bmatrix}$$

The synthesis of the LPV observer gains L_i for each vertex is achieved with LMI Toolbox of Matlab. The observer gain L of the system described by (58) is determined offline using (9) in the different simulation case bellow. The parameters of simulation are fixed as $\sigma = 10^{10}, \Gamma = 120 \, and \, \mu = 0.2$.

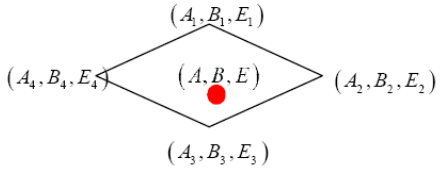
To illustrate the effectiveness of our algorithm, we choose a particular system matrix (A, B) computed using (4) for arbitraries values of $\theta_1 = 0.03$ and $\theta_2 = 0$:

$$A = \begin{bmatrix} -1.756 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ -1.8 & -1 & -0.75 & 0 \\ -1 & 0 & 0 & -1.006 \end{bmatrix} \quad (59)$$

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 0.494 \\ 1 & 0 \\ -0.006 & 0 \end{bmatrix} \quad (60)$$

$$E = \begin{bmatrix} 0 \\ 0.6 \\ 0 \\ 1 \end{bmatrix} \quad (61)$$

The given LPV system is defined in the vertex as shown in the figure above:



A. Case of constant fault signal

In the first case, consider the constant fault described by the following equation:

$$f(t)=\begin{cases}0 & t>0\\ 2 & 51\le t\le 20s\\ 0 & 101\le t\le 150\\ 2 & other\end{cases} \quad (62)$$

Figure 1 and Figure 2 show the estimation of the fault applied to the system described by matrix (59), (60) and (61).

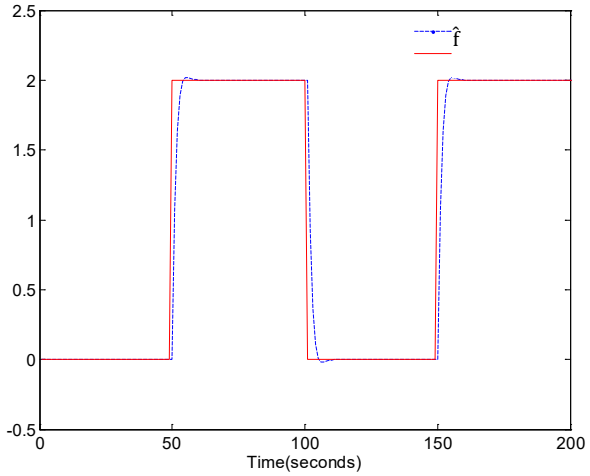


Fig. 1. f and its estimation $\hat{f}$.

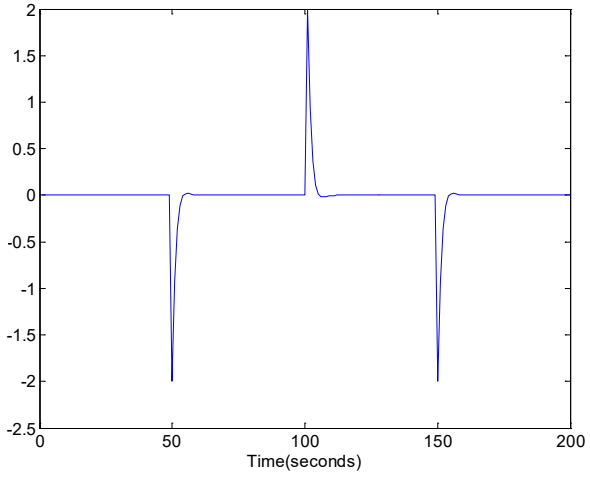


Fig. 2. The error $e_f(t)$ of the fault estimation

B. Case of variable fault signal

The actuator fault is described by:

$$f_{a2}(t)=\begin{cases}0.1\sin 5t+0.04\cos 3t\\ \quad +0.06\sin t+0.05\\ 0 \quad other\end{cases} \quad 7s\le t\le 12s \quad (63)$$

For the arbitrary values of $\rho(\theta)$, the simulation results are as the following:

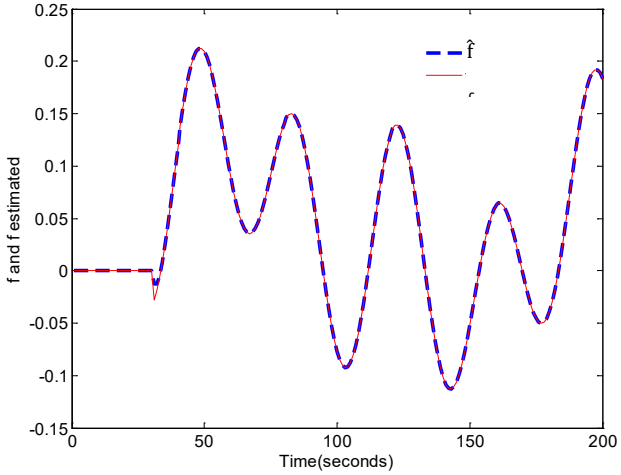


Fig. 3. f and its estimation $\hat{f}$.

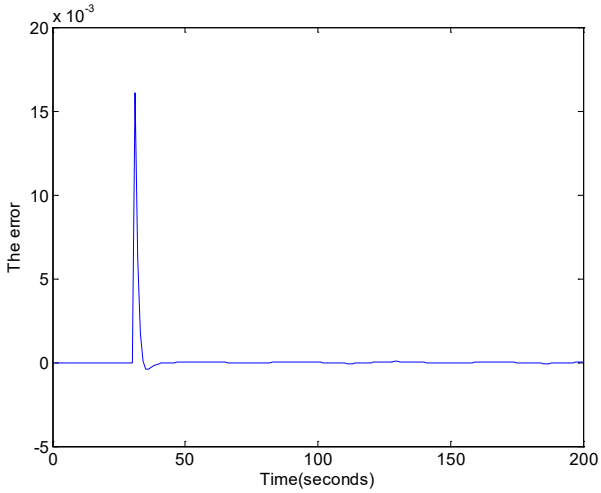


Fig. 4. The error $e_f(t)$ of the fault estimation

For the different types of fault signals considered above, Figure 2 and Figure 4 show the evolution of the error estimation. As can be observed, the error estimation converges asymptotic to zero even in the presence of disturbances. The real and the estimated constant and variable faults are displayed in Figure 1 and Figure 3 respectively. Summarizing, this approach can estimate the states and the fault functions with good performance and small error.

6. Conclusion

In this paper, an adaptive LPV observer using LPV approach has been presented for a polytopic system. A constant fault and a variable fault were considered. A polyquadratic Lyapunov function was used to perform the stability analysis. The problem was formulated in terms of linear matrix inequalities to develop the observer. The simulation results show the performances of the proposed observer. The main advantage of this representation is that it doesn't depend directly on the varying parameter. Moreover, this representation is defined as a difference between two parameters that evolve in two known and defined polytopes.

References

[1] K. Zhang, B. Jiang and V. Cocquempot, "Adaptive Observer-based Fast Fault Estimation" International Journal of Control, Automation, and Systems, **6**(3), 320-326, 2008. https://doi.org/10.1109/ChiCC.2015.7260636

[2] M. A. Montes de Oca, D. Aydın, T. Stutzle, "An Incremental Particle Swarm for Large-Scale Continuous Optimization Problems: An Example of Tuning-in-the-loop (Re) Design of Optimization Algorithms". Soft Computing, **15**(11):2233–2255. https://doi.org/10.1007/s00500-010-0649-0

[3] S. Grenaille, D. Henry and A. Zolghadri, "A method for designing fault diagnosis filters for LPV polytopic systems Journal of Control Science and Engineering. Volume 2008 (2008) : 1–11. http://dx.doi.org/10.1155/2008/231697

[4] Changhui W, Zhiyuan L., "A LPV adaptive observer approach to calibrate MAF sensor map in diesel engine". In: 54th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Hangzhou, China, pp.1375–1380. Piscataway, NJ: IEEE, 2015. https://doi.org/10.1109/SICE.2015.7285416

[5] C. Hoffmann and H. Werner, "A Survey of Linear Parameter-Varying Control Applications Validated by Experiments or High-Fidelity Simulations", IEEE Transactions on Control Systems Technology, Volume: **23**, Issue: 2, pp. 416 – 433, 2015. https://doi.org/10.1109/TCST.2014.2327584

[6] F. Bruzelius. "Linear Parameter-Varying Systems an approach to gain scheduling". PhD thesis, Department Of Signals And Systems, Chalmers University of Technology, 2004.

[7] I. Szászi, B. Kulcsár, G. J. Balas and J. Bokor, "Design of FDI filter for an aircraft control system", Proceedings of the American Control Conference Anchorage, AK May 8-10, 2002. https://doi.org/10.1109/ACC.2002.1024596

[8] F. Bruzelius, "Linear Parameter-Varying Systems: an approach to gain scheduling", Phd thesis, University of Technology of Göteborg, February 2004.

[9] F. Bruzelius, S. Pettersson and C. Breitholtz, "Linear parameter varying descriptions of nonlinear systems", in Proceeding of the IEEE American Control Conference (ACC), pages 1374–1379, Boston, Massachusetts, July 2004.

[10] O. Sename, P. Gáspár and J. Bokor, Robust Control and Linear Parameter Varying Approaches: Application to Vehicle Dynamics. Library of Congress Control, Number: 2012955946. Springer-Verlag Berlin Heidelberg, 2013.

[11] W. Xie, "H2 gain scheduled state feedback for LPV system with new LMI formulation", IEEE Proc.-Control Theory Appl., **152(**6), November 2005.

[12] D. Ichalal and S. Mammar, "On Unknown Input Observers for LPV Systems", IEEE Transactions on Industrial Electronics, 2015. https://doi.org/10.1109/TIE.2015.2448055

[13] J. Geromel and P. Colaneri, "Robust stability of time varying polytopic systems", Systems and control letters, **55**:81–85. 2006. https://doi.org/10.1016/j.automatica.2006.08.024

[14] Y. Caoand Z. Lin, "A Descriptor System Approach to Robust Stability Analysis and Controller Synthesis", IEEE Transaction on automatic control, Vol. 49, No. 11, November, 2004. https://doi.org/10.1109/TAC.2004.837749

[15] S.P. Boyd, L.E Ghaoui, E. Feron and V. Balakrishnam, Linear Matrix Inequalities in System and Control Theory, the Society for Industrial and Applied Mathematics, 3600 University City Science Center, Philadelphia, Pennsylvania 19104-2688, vol. **15**, 1994.

[16] P. Apkarian and R.J. Adams. "Advanced gain-scheduling techniques for uncertain systems". IEEE Transactions on Automatic Control, **6**:21–32, 1998. https://doi.org/10.1109/87.654874

[17] M. Jungers, P.L.D. Peres, E.B. Castelan, E.R. De Pieri, and H. Abou-Kandil. "Nash strategy parameter dependent control for polytopic systems". In 3rd IFAC Symposium on Systems, Strcuture and Control, Brazil, 2007. https://doi.org/10.3182/20071017-3-BR-2923.00097

[18] R. C. L. F. Oliveira, V. F. Montagner, P. L. D. Peres, and P.-A.Bliman. "LMI relaxations for H∞ control of time-varying polytopic systems by means of parameter dependent quadratically stabilizing gains". In 3rd IFAC Symposium on System, Structure and Control, Foz do Iguassu, Brasil, 2007. https://doi.org/10.3182/20071017-3-BR-2923.00099

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 451-459 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Innovative Research on the Development of Game-based Tourism Information Services Using Component-based Software Engineering

Wei-Hsin Huang[*,1], Huei-Ming Chiao[1], Wei-Hsin Huang[2]

[1]*Graduate Institute of Design Science, Tatung University & Department of Digital Game and Animation Design Taipei University of Maritime Technology, Taipei, Taiwan, R.O.C.*

[2]*Department of Media Design, Tatung University, Taipei, Taiwan, R.O.C.*

ARTICLE INFO

Article history:
Received: 14 December, 2017
Accepted: 18 January, 2018
Online: 10 February, 2018

Keywords:
Digital Guide
Game-Based Learning
Component-based Software Engineering

ABSTRACT

In recent years, a number of studies have been conducted exploring the potential of digital tour guides, that is, multimedia components (e.g., 2D graphic, 3D models, and sound effects) that can be integrated into digital storytelling with location-based services. This study uses component-based software engineering to develop the content of game-based tourism information services. The results of this study are combined with 3D VR/AR technology to implement the digital 2D/3D interactive tour guide and show all the attractions' information on a service platform for the gamification of cultural tourism. Nine kinds of game templates have been built in the component module. Five locations have completed indoor or external 3D VR real scenes and provide online visitors with a virtual tour of the indoor or outdoor attractions. The AR interactive work has three logos. The interactive digital guide includes animated tour guides, interactive guided tours, directions and interactive guides. Based on the usage analysis of the component databases built by this study, VR game types are suited to object-oriented game templates, such as the puzzle game template and the treasure hunt game template. Background music is the database component required for each game. The icons and cue tones are the most commonly used components in 2D graphics and sound effects, but the icons are gathered in different directions to approximate the shape of the component to be consistent. This study built a vivid story of a scene tour for online visitors to enhance the interactive digital guide. However, the developer can rapidly build new digital guides by rearranging the components of the modules to shorten the development time by taking advantage of the usage frequency of various databases that have been built by this study to effectively continue to build and expand the database components. Therefore, more game-based digital tour guides can be created to make better defined high-quality heritage attractions understood.

1. Introduction

In the United States, museums and historical sites generated 15 billion U.S. dollars in revenue in 2017 [1]. With the rapid development in information technology, the diversification and popularization of mobile carriers have made for rich digital content development. The Taiwan Tourism Bureau survey analysis also shows that cultural tourism has become an important tourism trend. The integration of multi-material, cross-platform and cross-industry values has become another focus of the content industry. Digital guides and online tourism maps of museums and historical sites' interactive tourist information services have gradually changed reading habits in recent years, replacing the hard copy of tourism guide maps. Since the development of the global digital economy, the output of digital content industries in Taiwan [2, 3] has reached over a 10% annual growth rate in recent years and a value of about 32 billion U.S. dollars in 2015. Based on the advantage of many excellent manufactories of information and communication technology in Taiwan, virtual reality technology has been regarded as an important driver of the digital economy. Applications such as commercial games, knowledge learning, and museum guides have increasingly introduced virtual reality, augmented reality, and mixed reality in Taiwan. For example, HTC and Foxconn Technology Group announced the first

*Corresponding Author: Wei-Hsin Huang, Tatung University, Taipei, Taiwan, R.O.C., Email: wshuang@ttu.edu.tw

VIVELAND™, which includes many virtual reality games, in Taiwan [4]. The Institute of Transportation adopted an interactive mixed reality platform [5] to develop a tool to assist teachers to teach children to walk safely around their schools [6]. The Shihsanhang Museum of Archaeology invites visitors to use a virtual reality tool to explore an important Taiwanese underwater cultural asset, the British steamship SS Bokhara, which sank near Taiwan in 1982 [7].

1.1. Digital Tour Guiding

Brown in [8] examined earlier applications that demonstrated location‐aware guiding by mobile devices, such as Cyberguide, which enabled a visitor to locate nearby bars, and GUIDE, which provided visitors with a means of personalizing information to suit their own interests and the environmental parameters. To increase the interactive elements, Bellotti in [9] built VeGame to operate on a cellular phone-mediated wireless connection between a pocket PC and the server, which communicated with the pocket PC via Bluetooth. VeGame was designed to enhance tourists' experience of art and history through a pleasant and challenging interaction with the heritage and people of Venice. When mobile devices became more widely used, Ballagas in [10] designed REXplorer, a device consisting of a Nokia mobile phone, a global positioning system (GPS) receiver, a camera, and a stretchable textile overlay with a zipper on the back that transformed a standard phone keypad into an 8-key game interface. Tourists could use REXplorer with gestures to interact with the significant buildings in Regensburg, Germany. Regarding indoor tour guiding, in addition to radio frequency identification (RFID) technology for positioning [11], Tsai in [12] proposed a position estimation method to design a location-aware tour guide system to be used with a personal digital assistant (PDA) by visitors to the National Palace Museum of Taiwan. As the user moved closer to items in the collection, the size of the materials became larger on the screen. Steiniger and Edwardes [13] indicated that location-based services (LBS) combined with GPS, geographic information systems (GIS), quick response (QR) codes, mobile facilities and scene-point tour databases are an important future direction of such technology. Currently, advances in mobile technology make it feasible to use virtual and augmented reality technology for learning [14]. Doong in [15] investigated the chocolate-related knowledge learning effect of a game-based learning system developed with a cross-platform of LBS and mixed reality technologies. The virtual treasure-hunt game world was the real location of a chocolate factory that the participants had to find on Google Maps.

1.2. Agile Development and Component-Based Software Engineering (CBSE)

Developing game-based digital content is an expensive and risky activity. Using CBSE methods allows one to design digital content quickly by reusing existing functions and models to produce a system [16]. Folmer and Mehm's overview of developing games with components presents a reference architecture that outlines the relevant areas of reuse and signifies some of the problems with developing components unique to the domain of gaming [17, 18]. Elements characteristic of games (good competition, character development, improvement, inspiring and creative challenges, and activity) can easily be adapted to the needs arising from cultural settings. The results would be an increase in motivation and the development of positive relations with the cultural subject. Based on the spirit of the agile development method, Wu Junyue (2013) devised a set of development processes for mobile customer service that are best for the mobile service application [19]. Wu Junyue's development process f reduces the output of unnecessary documents and uses repetitive development processes to shorten the development schedule. For these reasons, a challenge for the scientific community is to find new ways to visualize and disclose 3D digital contents to achieve better access to and communication of cultural heritage information [20]. Therefore, the purpose of this study is to determine how to develop digital guides rapidly and make them suitable for more kinds of handheld equipment.

1.3. 3D Virtual Game-based Environment

Effective learning is situated, active, and problem-based and requires immediate feedback [21]. Gaming elements provide an instructional environment with stimulation and learning motivation [22, 23]. A well-designed virtual educational game provides complex holistic problem-based environments, making it possible to develop situated understanding [24] and thus supporting effective learning. Virtual game-based environment designs follow the rules of game design. Nevertheless, the purpose is not only to entertain [25] but also to use the characteristics of video and computer games to create attractive and immersive learning experiences to achieve specific learning objectives [26]. In addition to the four important features (gameplay, feedback, interface, and challenge) of a good game [25], the design of a game-based learning system must further consider the realism of the game, the opportunities to explore and obtain new information, and the meaning of learning controls [26]

1.4. Agile Development and Digital Content Development

The effectiveness of the selection of the game template components and the game design is also the most thought-out part of the process of implementation. For example, Teng Feng Fishball Museum (TFFM) at Tamsui is a food tourism location where the TFFM combines the history and cultural characteristics of the Tamsui. The TFFM is focused on individuals or family members and is supported by a rich history and cultural knowledge. The TFFM can present more profound educational material and understanding with 3D VR game types. However, this kind of innovative food museum can also present a variety of innovative food products by focusing on entertaining 2D game templates. Both digital guides focus on increasing the pleasure of the user experience, allowing the user to learn more about tourist attractions, or increasing tourists' pleasure in traveling on their own. Digital guides also meet educational needs and draw visitors into visiting the museum.

2. Methods

2.1. The Game-based Tourism Information Service System (GTISS)

The Game-based Tourism Information Service System (GTISS) was developed for the purpose of breaking geographical and temporal restrictions. 3D virtual and game-based tourist attractions were created using Unity software. Three major elements are including in the GTISS: itinerary planner, games and

virtual reality design, and cultural tourism features. The concept of CBSE was adopted in the development of the GTISS by agile methods.

The GTISS is designed to break geographical and temporal restrictions by creating 3D virtual and game-based tourist attractions using Unity software. There are three major elements in the GTISS, including an itinerary planner, virtual game-based design, and cultural tourism features. The concept of CBSE was adopted in the development of the GTISS by agile methods. Employing CBSE in the design of a tourism information platform expedites the process, that is, multimedia components (e.g., game plans, 2D icons and pictures, music, sound effects, 3D models, and programs) in a previously created database can be used and reused for various projects, which lowers the threshold for developing a tourism information platform and adds value to existing tourism materials. Cultural tourism is often full of story-like elements; thus, to enhance the attractiveness of a cultural tourism information platform, it is necessary to improve its storytelling ability and incorporate the LBS and game elements into the platform [27]. Web3D, GIS, GPS, and Unity3D VR software were used in the present study. The goal of CBSE is to establish a digital content development platform consisting of reusable components; this platform can shorten the time for developing a project. Drawing on the concept of CBSE, the multimedia component database allows developers who do not have enough game development background to design the content of the platform easily. Finally, all the components in the database were created to be compatible with the Unity 3D engine.

To develop the GTISS, first, information content about tourist attractions was collected. Then, based on the features and characteristics of each tourist attraction and the corresponding game plans, appropriate components in the multimedia component database were searched to design a game that presents the information about each tourist attraction. LBS integrated the QR code to enhance the contextualized interaction. Tiered with levels of difficulty, it features requirements, such as time limits and penalties, that add excitement and competition to the game. The accomplishment of finding a destination involves answering questions or solving problems in some well-known scenic spots. Moreover, for users, the platform provides travel guides to the tourist attractions by showing available public transport through dynamic maps and consecutive pictures of real street views. In addition, by using Google Maps application program interfaces (APIs), the platform shows the GPS locations of the tourist attractions on Google Maps, which is familiar to many users.

2.2. Agile Development and Digital Content Development

A number of tourist attractions that the government actively promotes were selected for information collection. Each digital guide for the tourist attractions provides public transport information integrated into the digital map to promote a Low Carbon Tour. Applying the appropriate game plan for an attraction's information confirms that the characteristics of the attraction's content are sufficient for the resources required for the game plan. For example, if there is not enough information about the attraction, then the use of the real tour will lead to digital guide content that is not attractive. Finally, we applied the modular component and game design principles of the plan to archive the digital guides. With the Unity 3D engine, the completed digital guide can be built into different carrier formats (such as html, apk, and exe). The process of component-based agile game and digital content development used by this study is shown in Figure 1.

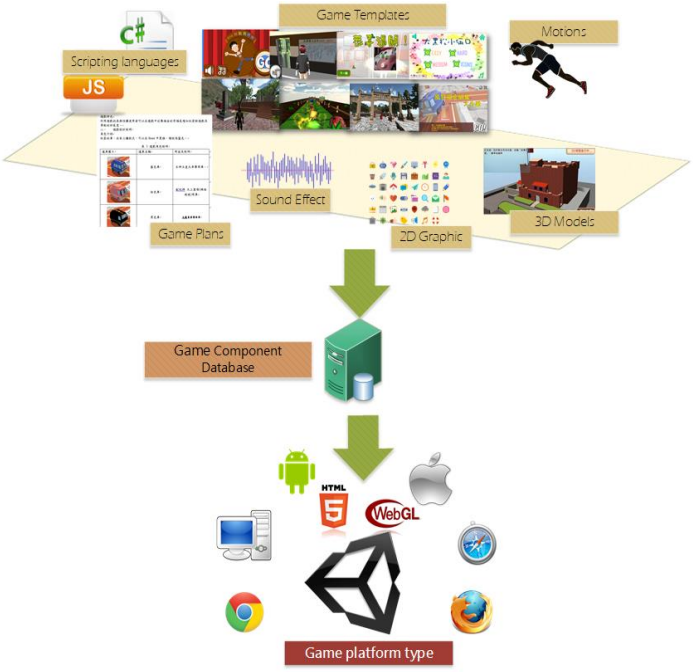


Figure 1 The process of use agile design method to make digital content

The effectiveness of the selection of the game template components and the game design is also the most thought-out part of the process of implementation. For example, Teng Feng Fishball Museum (TFFM) at Tamsui is a food tourism location where the TFFM combines the history and cultural characteristics of the Tamsui. The TFFM is focused on individuals or family members and is supported by a rich history and cultural knowledge. The TFFM can present more profound educational material and understanding with 3D VR game types. However, this kind of innovative food museum can also present a variety of innovative food products by focusing on entertaining 2D game templates. Both digital guides focus on increasing the pleasure of the user experience, allowing the user to learn more about tourist attractions, or increasing tourists' pleasure in traveling on their own. Digital guides also meet educational needs and draw visitors into visiting the museum.

To rapidly develop a game-based digital guide, component-based templates were used to develop a digital guide of tourist attractions. The construction of the modular component database replaced the homemade or commercially available development resource modularization database. Seven component databases were established and used to develop a game-based digital guide that can be viewed on webpages or handheld devices. The component databases are shown in Table 1. A motion database and a game template database were recently completed. After developing several games, the humanoid character can be used in a cross-game motion template. We expanded our motion database

by using x-box to scan simple motions that the game needed and binding the motion to the characters.

Table 1: Component-Database of this study

Database Name	Content
3D Model	The 3D model library contains models such as character, buildings, and objects that built from a variety of modeling software, such as 3D Max, Maya, and SketchUp. All models are finally imported in the development engine with .fbx format
2D Graphic	This database contains illustrations for animations and interface designs such as picture or icon. The 2D material in database licensed under creative commons. Some of these art files may contain .jpg, .png, and so on in different file formats.
Scripting language	This study aimed at scripting and sorting all kinds of shareable functions in game templates. For example, the function of these feet for the game interface to return and explain the function, the game time count/pause/stop, props configuration, character action control. All scripts has been documentation for re-use by non-game developers.
Game Plan	The game template of this study will be self-development and collect from web resource database. We write planning documents for each free or paid game template for non-game developer to use.
Sound Effect	Collect, organize and use the sound studio or online recording software produced by the various sound effects database. All content licensed under creative commons, and contains include .mp3 or .wmv and other file formats of sound.
Motion	In addition to collecting free motion files available on the network, find out how to organize the motion files from ready-made game templates into other game templates. This study also used X-Box for Motion capture to record the movements of real people and mapping the movements onto the CG character.
Game Template	By implement the game plan into real game, there are eight game template had been build. There are nine game template in our database.

In addition to the integration of models, animations and sound effects in the game development process, the development of game script language is also very important. However, the number of script resources that can be used directly in the game is less than the number of models or sound effects. As shown in Figure 2, if building a script language database can be organized, then the time that programmer used to rewrite the code can be saved. The developer simply confirms the required functionality, finds the appropriate script resource in the database, and imports it into the Unity game engine. Fine-tuning allows game developers to quickly develop game features and add new script into the script database.

2.3. 360-Degree Virtual Tour Guide

To increase the sense of participation, an attempt was made to imitate the indoor and outdoor feeling of the attractions that visitors would experience. Some of the attractions are not well defined by their location or are connected to a factory, and the user cannot find the entry.

This study uses Google Street View's screenshots to build a 360-degree VR scene. The application of this 360-degree scene setting in this platform mainly provides an indication of the user's arrival status at the attraction. Using the 360-degree circular guide at the entrance increases the user's understanding of the geographical location of the scenery. The design of this scene can be applied to the design of future AR scenes. The scene of this study is based on the use of Google Street View and real indoor photos made during field trips to build outdoor and indoor scenes. In the future, the user may use the camera lens of the mobile device as a source, and the digital guide instructs the user to provide information or tips based on all the current locations.

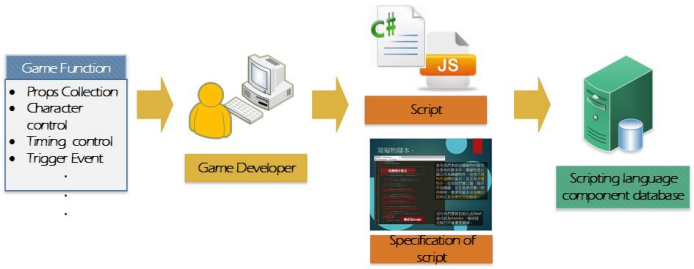


Figure 2 The process of build Scripting language component database

3. Results

3.1. Game-based Component Database

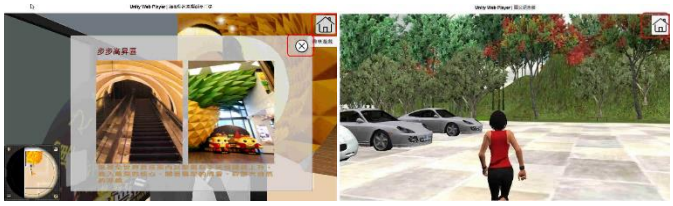
This study built seven component databases, as described in Table 1. Figure 3, Figure 4 and Table 2 are the samples of the 3D models, different motions and sound effects collected in this study in the present component databases. Regarding the 3D character models, for example, there are currently ten 3D character models in the database. As shown in Figure 3, the same character can be used in different games quickly with various motions. The movement of running includes running forward, turning left, and turning right. The motions of touching the head and waiting are bound to form the other motion. Some motion files for men and female are different. In the different games shown in Figure 3, it is possible not only to reuse the 3D character model in the game template but also to use 3D models of terrain and trees. User interface (UI) templates in the 2D graphic database can be created for different attractions easily, quickly and repeatedly. As shown in Figure 4(a), the same UI template was used in the National Dr. Sun Yat-sen Memorial Hall and the Vigor Kobo Dream Museum. Through a color change, the same UI component can quickly adapt to each game template. Similarly, by using consistent icons, the user can understand the meaning of the graphic on the screen. As shown in Figure 4 (b), the icons on the screen are the indispensable elements that guide the user to operate the icon related functions. During the preliminary stage of design, it took about a month to accomplish the construction of a scenic spot whereas only five working days were needed in the later stage for the creation and modification of an attraction, which is the most time-consuming of the entire project.

(a) Turn-left running motion on female (b) Turn-left running motion on man

(c) The waiting motion (d) Touch head motion and two character

Figure 3 The 3D character models and motions in different game

(a) The UI template used by two attraction



(b) The icon used by different game

Figure 4 The 2D UI template and close icon be reuse in our project for saving the time of development

3.2. The Tourism Digital Guiding Platform (TDGP)

In the present study, the GTISS was designed and shown as in Figure 5. All attraction locations were labeled in Google Maps and categorized into six types: food, shopping, heritage, transportation, education, and entertainment. User-friendly interfaces enable visitors to find interesting locations easily. Fifteen cultural locations were chosen and designed in the TDGP. Figure 6 shows the template for planning the route to the destination (e.g., from Taipei Main Station to the Taiwan Socks Museum). Users can choose to board at the mass rapid transport (MRT) station. Then, according to the routing algorithm, the user can achieve the minimum number of transfers or the shortest distance calculated from the number of stops. Both the interactive display technology of the LBS geographic location and the 3D VR/AR are important trends in the digital content industry.

The interactive traffic information guideline was provided by the destination websites. Combined with the street/attractions and

Table 2 Component-Database of this study

Attractions	BGM
Anping Castle	Path to Follow.mp3
Republic of Chocolate	The_Messenger.mp3
National Dr. Sun Yat-sen Memorial Hall	Clouds.mp3
Longshan Temple	Clouds.mp3
Taiwan's Socks Museum	On the Bach.mp3
W & W Museum of Jewelry	On the Bach.mp3
Vigor Kobo Dream Museum	Festival.mp3
National Palace Museum	The_Engagement.mp3
Lanyang Museum	Retreat.mp3
Dali Tian Gong Temple	Succotash.mp3
Kuo Yuan Ye Museum of Cake and Pastry	Reasons_to_Smile.mp3
Taiwan Nougat Creativity Museum	Morning Walk.mp3
KONG YEN Yi Shou Dou Cultural Museum	Microchip.mp3
Yumeeriren maternity tour factory	Keith.mp3
Shu Shin Bou Wagashi Museum	Morning Walk.mp3

Figure 5 The tour platform use Google Maps API to presentation

Figure 6 Dynamic route planning to Taiwan's Socks Museum(

outdoor/indoor navigation technology of an attraction's features, users can have an in-depth understanding of the tourist information and be impressed by the digital tour guides.

The completion of each level involves answering questions and solving problems. Answers to the questions can only be found by asking locals questions, exploring the surroundings, or using certain services specific to a location. At the outset, the QR code of a zone must be photographed, and when all the questions in that zone are answered or when the problems are solved, the user can then advance to the next level of his/her choice.

Figure 6 shows useful travel guides included in the tourist attractions with available public transport route information through dynamic maps. By using the 360-degree circular guide, consecutive pictures of real street views can be provided to increase the visualization and understanding of those locations (shown in Figure 7). The platform shows the GPS locations of the tourist attractions on Google Maps, and more LBS information can be added to the system.

The 2D/3D game-based learning content of fifteen attractions are shown in Figure 8. These are famous attractions in Taiwan and are promoted by the Taiwan Tourism Bureau. The development of a digital tour guide is mainly based on 3D game types. However, the architecture is the main part of building the 3D models of the attractions, and it requires more time and work to adjust during the development process. To attain the purpose of the agile development of the new digital guide, this study also built some interesting 2D game templates for our database to save time and reduce the needed process in developing digital tour guides, as shown in Figure 8.

3.3. 3D AR / VR Game-based Tourism Service

Seven databases were used in designing the 15 attractions for the effective and efficient development of the game. Table 3 shows the number of 2D graphics with the game templates applied in the 15 attractions. Figure 9 shows several 2D graphics used in the game templates. Most reusable images are icon-type graphics, which can change color, size, and text. In addition, each game has background music, which was chosen from the background music (BGM) database based on rhythm and types. Table 4 shows all the music applied in the game templates for the study.

To create a 3D VR game template, a goal-oriented puzzle or treasure hunt was created by interacting with the object in the scene (Table 5). The Unity3D game engine was used to build indoor and outside VR panoramas of attractions in this tourism platform. Google street view was used for outdoor scenes, and 3D models with real indoor photos of the museums were created in the present study. Through the implementation of an online VR tour, users can have pleasure and further understand the details of the attractions.

There are nine kinds of game templates built in the component modules. By using these component modules, a digital guide with 3D VR real scenes were provided to visitors for touring both the indoor and outdoor attractions (as is shown in Figure 10). This study organized the game-based component templates and created vivid scenarios of scene tours for visitors to enhance the interaction with the digital guide. Nine kinds of game templates with three game types are shown in Table 5.

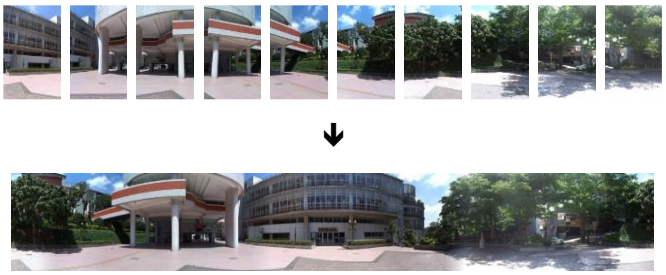
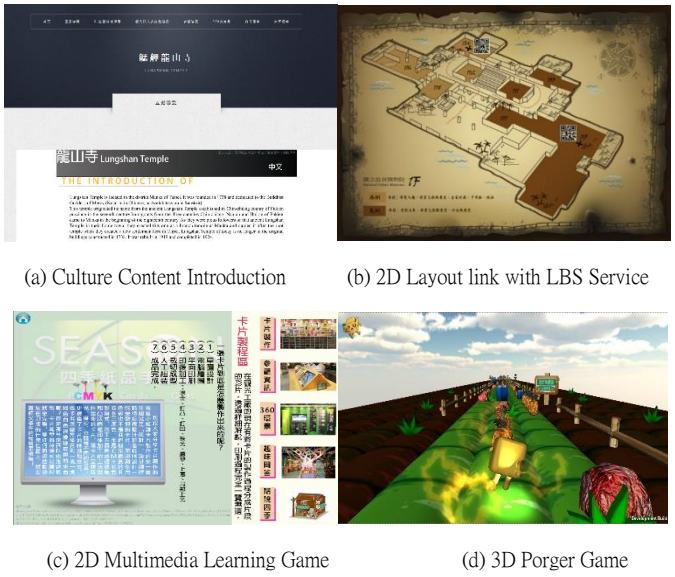


Figure 7 This study use Google Street View's screenshots to build up a 360-degree VR scene



(a) Culture Content Introduction (b) 2D Layout link with LBS Service

(c) 2D Multimedia Learning Game (d) 3D Porger Game

Figure 8 The Cultural Tourism Digital Guiding Platform

Table 3 Number of Sound used by 15 Attractions with Game type

Sound	Game Type			Total
	2D	3D	3D VR	
BGM	**5**	**5**	**5**	**15**
Bird		1		1
Cannon sound			1	1
Cue Tone -Button Click	4			4
Cue Tone -Notice		1	4	5
Cue Tone -Pass	1	1		2
Cue Tone -Pick up Item	1	2		3
Cue Tone -Wrong	**4**	**1**	**4**	**9**
Cue Tone-Car Crash		1		1
Cue Tone-Loss Health points		2		2
Cue Tone-Miss Hit	1			1
Cue Tone-Most Time Up	4			4
Cue Tone-Notice		1		1
Cue Tone-Right	**4**	**1**	**4**	**9**
Cue Tone-Time Up	4			4
Explosion		1		1
Mining sound			1	1
Racing Sound		1		1
Shooting sound	1	1		2
Train arrive		1		1
Typing Effect		1		1
Water of fountain		1		1
Wheel Rolling Sound	1			1
Wind sound			1	1
Total	30	22	20	72

Table 4 Number of Attractions with Game template and Game type

Game Template	Game Type			Number of Attractions
	2D	3D	3D VR	
Book Template		1		1
t Wheel Game Template	1			1
t Personal Shooting Game		1		1
den Object Game Template	2			2
stery Jigsaw Game Template	2			2
kour games Template		1		1
zle Game Template			2	2
ing Game Template		1		1
asure hunt game Template		1	3	4
al	5	5	5	15

Table 5 Number of 2D Graphic used by 15 Attractions with Game type

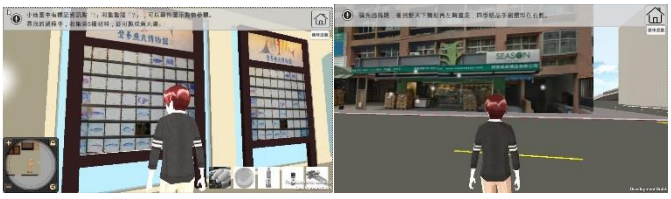
2D Graphic	2D			2D Total	3D					3D Total	3D VR	
	A.	B.	C.		D.	E.	F.	G.	H.	I.		J.	K
Entry Graphic	1		2	3		1					1		
Exit Icon	**1**	**2**	**2**	**5**			**1**	**1**	**1**		**3**		
Game Over	**1**	**2**	**2**	**5**			**1**	**1**	**1**		**3**		
Home Icon	**1**	**2**	**2**	**5**		**1**			**1**	**1**	**3**	**2**	
Level-Hard		2		2									
Level-Midium		2		2									
Level-Simple		2		2									
Man-Waiter	1			1									
Next icon			2	2									
Next-Page						1					1		
Pause								1			1		
Pre-Page						1					1		
Replay Icon	**1**	**2**	**2**	**5**			**1**	**1**	**1**		**3**		
Right						1				1	2	2	
Setting Icon	1	2		3					1		1		
Store									1		1		
Success								1		1	2	2	
Treasure Map													
Try Again										1	1		
UI Background	1	2	2	5									
Wrong						1					1	2	
總計	8	18	14	40		6	3	5	6	4	24	8	

Notice the Symbol of Game Template :

A. Dart Wheel Game Template
B. Hidden Object Game Template
C. Mystery Jigsaw Game Template
D. 3D Book Templete
E. First Personal Shooting Game
F. Parkour games Template
G. Racing Game Template
H. Treasure hunt game Template
I. Puzzle Game Template
J. Treasure hunt game Template

Figure 9 The 2D Gameful Design Digital Guide of this research

(a) The 3D Geometry-base Model VR Digital Guide



(b) The Image-base Background with 3D Geometry-base Model VR Digital Guide

Figure 10 The Digital Guide of this research

During the 2D game development period, a navigation e-book was also built to be merged with the 3D model. The design of the navigation e-book has been improved from the original 2D material to the 3D model library by adding a UI interface for a more interesting and clear introduction. To upgrade the function of the navigation e-book content, we incorporated 2D text descriptions and dynamic 3D guidance to enhance the entire navigation e-book'tals readability. This navigation e-book is released on a Unity platform. There is a PC version, a web format, and an .apk format for other handheld devices.

Animations and interactive elements were added to this navigation e-book to increase the interactivity between the digital content and the users. Using Lan Yang Museum's 3D navigation e-book as an example, on the floor map page, by clicking on the 3D text on the left and the red indicator on the right side of the page, it will move to the floor on the map (Figure 11(a)). In the traffic guide page, the 3D models of moving trains and buses, as shown in Figure 11(b), were also added to improve the appeal of the e-book as well as the enjoyment of the visitors.

To further expand our VR games into AR games. Three museums' logo images were chosen as the AR image-based

(a) PC version 2D e-book model (b) 2D e-book with UI

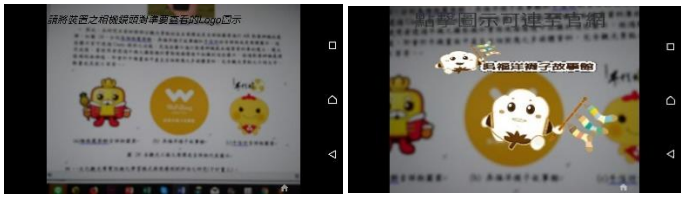
(c) E-book with 2D map and 3D words (d) E-book With dynamic 3D model (Bus and Train)

Figure 11 The screenshots of Lan Yang Museum's navigation e-book (apk)

identification, which can generate four identifications supported by the Unity software: single picture identification, cuboid identification, cylinder identification and 3D object identification. When the AR glass lens captures the correct image object, other instructions from the weblink or hyperlink to the museum website are triggered (as shown in Figure 12).

4. Discussion

(a) The screenshot of mobile device that hint user how to use AR

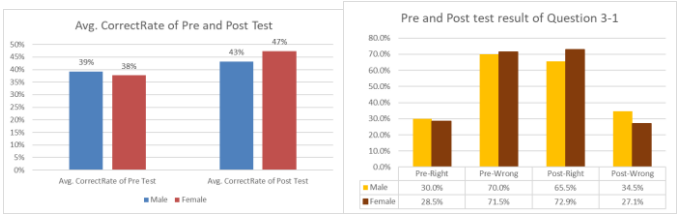


(b) The logo photo capture by mobile device (c) Show the Mascot's Animation

Figure 12 Screenshot of AR identification in this research

It is important to design and develop game-based digital tour guides using agile development and component databases in the future. Game-based digital guide learning easily increases not only the entertainment effect but also the understanding and recollection of the cultural attractions. However, to accomplish this complex work (drawing the 3D model, writing the coding, producing the music, etc.) requires a long development process. Developers with experience in game design need at least a month to finish a game project without using a component database. Using the game-based component database, intern students with no experience in game development were trained to complete the first digital game guide within two weeks. After completing the first project, four intern students found new attractions and used game templates to develop new game-based digital guides on their own. The intern students can not only develop independently but also shorten the design period to within a week. In the future, non-game professional tour staff can complete their own interesting digital tour guides easily and quickly.

After the platform is established, the attention of younger users can be captured by measuring the accuracy of the questions that users answer before and after using the platform. If the digital tour guide of this platform is sufficient to attract users to browse, read and memorize the content, the accuracy of the answers in post-test results should be improved. There were 477 student participants. Their ages ranged from 18 to 22 years old, and the number of female students (338 or 70.9%) was higher than the number of male students (139 or 29.1%). This study used a pre-test and post-test on knowledge content design issues for the platform attractions. The pre-and post test aimed to examine whether students enhanced their understanding of the attractions and the online travel planning. The items include an itinerary planner, games and virtual reality design, cultural features of the attractions, and Tourism English. Students did the pre-test before the experiment and the post test after it. The experiment lasted for one hour. The purpose of the system design was explained and the operational functions were demonstrated before the experiment. Pre- and post-tests were created based on the content of the system to examine student learning on the platform. As shown in Figure 13 (a), both males and females increased their correct answer rate after using the platform. Figure 13 (b) shows the pre- and post-test correct answer rates for question 3-1 in the questionnaire. The post-test correct answer rates for both males and females were 10% or higher than the pre-test. These results are sufficient to initially demonstrate the validity of the digital tourism content on the platform established by this study. The game-based digital tour guides with locality-specific information and culture in mind can make not only real hands-on experiences but also experiences that are memorable and worth sharing.



(a) Avg. CorrectRate of Pre and post test (b) Pre and Post test result of Question 3-1

Figure 13 Two result of pre and post test

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is supported by the Ministry of Science and Technology, Taiwan, R.O.C. under Grant no. MOST 104-2632-H-262 -001 -MY2.

This paper was originally published in the 2017 IEEE International Conference on Applied System Innovation(ICASI), May 13-17, 2017, Hotel Emisia, Sapporo, Japan.

References

[1] Courtney Tucci. Statistics and Facts on Museums. Retrieved from https://www.statista.com/topics/1509/museums/ (2015).
[2] [2] Ministry of Culture (2015). Taiwan Cultural Creative Industries Annual Report 2016 (in Chinese). Ministry of Culture.
[3] [3] Institute for Information Industry (2016). 2015 Digital Content Industry in Taiwan (in Chinese). Ministry of Economic Affairs, Industrial Development Bureau.
[4] [4] HTC. VIVELAND (https://www.vive.com/tw/viveland/).
[5] [5] Ji-Liang Doong Ching-Huei Lai, Kai-Hsiang Chuang, Chun-Chia Hsu (2015, Jul). The Development of Location-based Mobile Augmented Reality (LoMAR). The 3rd International Conference on Hospitality, Leisure, Sports, and Tourism. July 23-24, 2015, Waseda University, Tokyo, Japan .
[6] [6] Chun-Chia Hsu, Chih-Yung Lin, Yu-Li Chen, Wei-Shin Huang, Kai-Hsiang Chuang, Huei-Ming Chiao in (2017) Tools development of investigating roads and using roads for pedestrians and bicyclists (1/3) – Applications of road safety audit (in Chinese). Institute of Transportation, Department of Transportation and Communications.
[7] [7] The Shihsanhang Museum of Archaeology. Virtual reality exploring the S. S. Bokhara (http://www.sshm.ntpc.gov.tw).
[8] [8] Brown, Elizabeth; Brner, Dirk; Sharples, Mike; Glahn, Christian; de Jong, Tim and Specht, Marcus (2010). Location-based and contextual mobile learning. A STELLAR Small-Scale Study. STELLAR European Network of Excellence in TEL (EU).
[9] Bellotti, F., Berta, R., De Gloria, A., Ferretti, E., & Margarone, M. (2003). VeGame: exploring art and history in Venice. Computer 36(9), 48-55
[10] Ballagas, R., Kuntze, A. and Walz, S. P. Gaming Tourism: Lessons from Evaluating REXplorer - a Pervasive Game for Tourists. In: J. Indulska in (Eds.): Pervasive 2008, LNCS 5013, pp. 244–261, 2008. Springer.
[11] Wang, Y., Yang, D., Liu, S., Wang, R. and Meng, X. A RFID & Handheld Device-Based Museum Guide System. in Proceedings of the International Conference on Pervasive Computing and Applications (ICPCA), 2007, pp.308–313.
[12] Tsai C Y, Chou S Y, Lin S W. Location-aware tour guide systems in museum. Scientific Research and Essays, 2010, 5(8): 714-720.
[13] Steiniger, S., Neun, M., & Edwardes, A. (2006). Foundations of Location Based Services: University of Zurich.
[14] Specht, M., Ternier, S., & Greller, W. (2011). Dimensions of mobile augmented reality for learning: a first inventory. Journal of the Research Center for Educational Technology, 7(1), 117–127.
[15] Ji-Liang Doong, Ching-Huei Lai, Kai-Hsiang Chuang, Chun-Chia Hsu, Learning Effects of Location Based Mixed Reality Game: A Pilot Study, Procedia Manufacturing, Volume 3, 2015, Pages 1603-1607.World Tourism Organization. (2017).UNWTO Tourism Highlights: 2017 Edition - UNWTO Elibrary
[16] Bei-lin Lin, “On the rebuild-of the XBRL Demosite by using component-based design.”, Master Thesis, National Central University, 2011
[17] Folmer, E., Component based game development : A solution to escalating costs and expanding deadlines. CBSE'07 Proceedings of the 10th international conference on Component-Based Software Engineering(pp. 66-73) (ISBN: 978-3-540-73550-2) (2007)
[18] Florian Mehm, Stefan Gobel, Ralf Steinmetz. Introducing component-based templates into a game authoring tool. In Dimitris Gouscos and Michalis Meimaria (ed.) 5th European Conference on Games Based Learning (pp. 395–403)., Reading, UK: Academic Conferences Limited. (2011)
[19] JYUN-YUE WU, "Applying Agile Method to Building Mobile Application Software and Services", Master thesis, Shih Hsin University,2013
[20] Poria Y, Bulter R, Airey D. The core of heritage tourism. Annals of Tourism Research, 2003, 30(1): 238-254.
[21] Boyle, E., Connolly, T. M., & Hainey, T. (2011). The role of psychology in under- standing the impact of computer games. Entertainment Computing, 2(2), 69–74.
[22] Din H W H. Play to learn: Exploring online educational games in museums. In The 33rd International Conference and Exhibition on Computer Graphics and Interaction Techniques, Boston, MA, 2006.
[23] Sørensen B H, Meyer B. Serious games in language learning and teaching - A theorectical perspective. In Proceedings of 3rd International Conference of the Digital Games Research Association, 2007, pp. 559-566.
[24] Annetta L A, Folta E, Klesath, M. V-Learning: Distance Education in the 21st Century through 3D Virtual Learning Environments. Springer, New York, 2010.
[25] Johnson, L.W., Vilhjalmsson, H., Marsella, S. in: Artificial Intelligence in Education, IOS Press, Amsterdam, Netherlands, 2005, pp. 306-313.
[26] de Freitas, S. Learning in immersive worlds, 2006. (http://www.jisc.ac.uk/media/documents/programmes/elearninginnova-tion/gamingreport _v3.pdf).
[27] Ghiani G, Paternò F, Santoro C, Spano L D. UbiCicero: A location-aware, multi-device museum guide. Interacting with Computers, 2009, 21(4): 288-303.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 460-466 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Performance Analysis of Regenerative Braking in Permanent Magnet Synchronous Motor Drives

Andrew Adib*, Rached Dhaouadi

College of Engineering, American University of Sharjah, P.O. Box 26666 Sharjah, UAE

ARTICLE INFO

Article history:
Received: 11 November, 2017
Accepted: 09 January, 2018
Online: 10 February, 2018

Keywords:
Permanent Magnet Synchronous Motor
Regenerative Braking
DC-DC Converter
Ultracapacitor
Pulse Width Modulation
Energy Harvesting

ABSTRACT

This paper describes the design and analysis of a regenerative braking system for a permanent magnet synchronous motor (PMSM) drive for electric vehicle (EV) applications. First studied is the principle for electric braking control of a PMSM motor under field-oriented control (FOC). Next, the maximum braking torque in the regeneration mode as well as the braking torque for the maximum regeneration power, respectively, are deduced. Additionally, an optimum switching scheme for the inverter is developed with the objective of maximizing energy recovery during regenerative braking to the DC-bus capacitor. The integration of an ultracapacitor module with the battery allows for the efficient and high power transfer under regenerative braking. It was important to manage the power flow to the DC-bus as this is a key issue that affects the efficiency of the overall system. Finally, the amounts of braking energy that can be recovered, and the efficiency with which it can be returned to the battery/ultracapacitor, is analyzed for a PMSM coupled with a DC motor as the load. The results of the analysis are validated through experimentation.

1. Introduction

Interest in regenerative braking is growing drastically nowadays; as the market is slowly transitioning to electric vehicles (EV) instead of the traditional vehicles that run on fossil fuels. Regenerative braking (RB) utilizes the kinetic energy generated by the motor during the deceleration, or braking, process. Therefore, recovering the braking energy is an effective approach for improving the driving range of an EV [1]-[4]. Usually, within traditional vehicles, all of the braking energy is lost in the form of heat due to friction losses. In RB, the motor acts as a generator and the kinetic energy is harvested by applying the proper switching schemes to the power converter switches. This harvested energy can be used to charge the vehicle's battery, or stored in an ultracapacitor bank [5]-[8]. A hybrid energy storage system can be used to alternate power generation and storage between an ultracapacitor and a battery, depending on the required power. Knowing the parameters of the system is essential to building a clear idea regarding the amount of energy harvested as opposed to that being generated.

The permanent-magnet synchronous motor (PMSM) is widely adopted as the traction motor in electric vehicles (EV) due to its high efficiency and high torque density. Vector control, also called field-oriented control (FOC), is a popular and powerful method in electrical drive applications. This control strategy is used to effectively control the PMSM motor torque and flux in order to force the motor to accurately track the command trajectory regardless of machine and load parameter variations, or any other external disturbances [9]. Electric braking control of the PMSM based on FOC is realized by requesting a negative q-axis current according to the braking torque demanded. The maximum amount of current produced while braking is calculated depending on many variables, including the motor speed and input voltage [10].

This paper is an extension of work originally presented in ICMSAO'17, where regenerative braking was analyzed for a DC motor with battery/supercapacitor energy storage [11]. The maximum amount of current produced by the DC motor while braking was calculated depending on the system variables such as motor speed, armature resistance, and input voltage. The theoretical analysis was next validated by experimental results. The effect of varying the duty cycle of the braking signal was also studied to find the optimal duty cycle that gives the best efficiency in regenerative energy harvesting.

In this paper, electric braking is first analyzed for the PMSM under Field Oriented Control (FOC). Next, a dedicated maximum Energy Recovery Switching Scheme (MERSS) is developed to control the inverter switches during regenerative braking to maximize energy recovery. Regenerative braking energy calculations are confirmed by experimental results on a prototype

*Corresponding Author: Andrew Adib, Email: b00062741@aus.edu

PMSM motor drive system. The harvested energy is stored into the DC-bus capacitor. The efficiency with which energy can be returned to the ultracapacitor is analyzed for the FOC scheme and the MERSS. The analysis, benefits, limitations and experimental results of each control strategy are provided.

This paper is divided into 5 sections; section 2 presents the concept of regenerative braking under FOC and provides some theoretical analysis of the required motor current during this mode of operation. Section 3 presents the maximum energy recovery switching scheme. Section 4 presents the experimental results and discussion when braking the PMSM under FOC and MERSS. Finally, the conclusion is presented in Section 5.

2. Regenerative Braking of the PMSM Under FOC

In conventional passive dynamic braking the kinetic energy of electric machines is dissipated through the armature coils and the additional braking resistors. The principle of the regenerative braking with the voltage-source inverter is similar to that of the conventional dynamic braking, however; the additional resistors are eliminated and no additional power switches are used.

According to the driving conditions, the braking process can be classified as either constant speed braking or variable speed braking. The constant speed braking is usually required during the downhill road driving, while the variable speed braking often occurs during the general deceleration process.

2.1. System Setup

Figure 1 shows the general topology used in electric vehicle applications with hybrid energy storage, where the mechanical load is typically coupled to a Permanent-Magnet Synchronous Machine supplied by a battery source through an inverter. An ultracapacitor module is used as an auxiliary power source connected through a bi-directional DC-DC converter to the DC-link, thus making it possible to obtain an optimized charge/discharge operation mode.

2.2. Regenerative Braking principle of the PMSM

The FOC scheme for the PMSM is shown in Fig. 2. In FOC, the stator phase currents are measured and converted into a complex vector. This current vector is then transformed to a coordinate system rotating with the rotor of the machine.

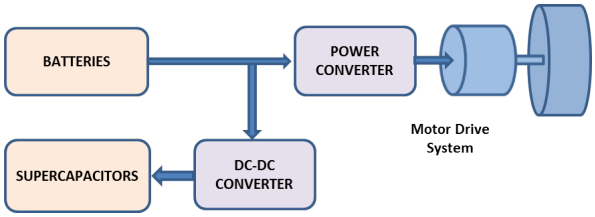


Figure 1: Topology of the Battery-Ultracapacitor Energy Storage System

The real x-axis component of the stator current vector i_d, in this rotor flux-oriented coordinate system, is used to control the rotor flux linkage. The imaginary y-axis component i_q is used to control the motor torque. For the PMSM, maximum torque–current control can be achieved by holding the d-axis current at zero ($i_d = 0$).

Electric braking control, based on FOC, is realized by requesting a negative q-axis current according to the braking torque demanded, or by controlling the speed to follow a ramp reference and gradually approach zero [6].

The input power and the electromagnetic power can be expressed as

$$P_{in} = V_d i_d + V_q i_q, \quad (1)$$

$$P_{em} = T_{em}\omega\,, \quad (2)$$

$$T_{em} = \frac{3}{2}\frac{p}{2}\lambda_{PM} i_q\,. \quad (3)$$

where, i_d and i_q are the d-axis and q-axis current components, V_d and V_q are the d-axis and q-axis voltage components, ω is the motor speed, T_{em} is the electromagnetic torque, λ_{PM} is the permanent magnet flux, and p is the number of poles.

The electric machine power losses are given as

$$P_{loss} = i_d^2 R + i_q^2 R. \quad (4)$$

where, R is the per-phase stator resistance. By balancing the input and output power

$$P_{in} = P_{em} + P_{loss}, \quad (5)$$

$$P_{in} = \frac{3}{2}\frac{p}{2}\lambda_{PM} i_q \omega + R(i_d^2 + i_q^2). \quad (6)$$

To find the regions of regenerative braking, P_{in} is set to 0. This means that no power is being drawn from the DC source.

Next, an expression for i_q in terms of i_d can be found

$$i_q = \frac{-\frac{3p}{22}\lambda_{PM}\omega \pm \sqrt{\left(\frac{3p}{22}\lambda_{PM}\omega\right)^2 - 4R^2 i_d^2}}{2R}. \quad (7)$$

The equivalent electromagnetic torque is given as

$$T_{em} = \frac{3}{2}\frac{p}{2}\lambda_{PM}\left(\frac{-\frac{3p}{22}\lambda_{PM}\omega \pm \sqrt{\left(\frac{3p}{22}\lambda_{PM}\omega\right)^2 - 4R^2 i_d^2}}{2R}\right). \quad (8)$$

This equation describes the electromagnetic torque inside the regenerative braking region. To find the boundaries of this region, the above equation is solved for the maximum and minimum braking torque.

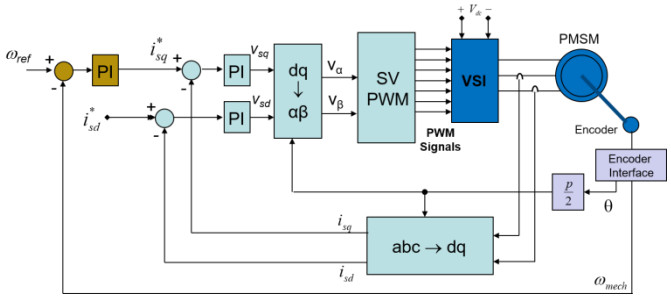


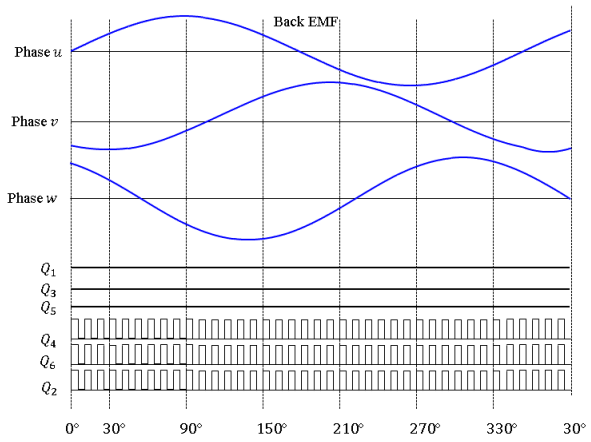
Figure 2: Field Oriented Control Scheme of the PMSM

$$\frac{dT_{em}}{di_d} = 0. \quad (9)$$

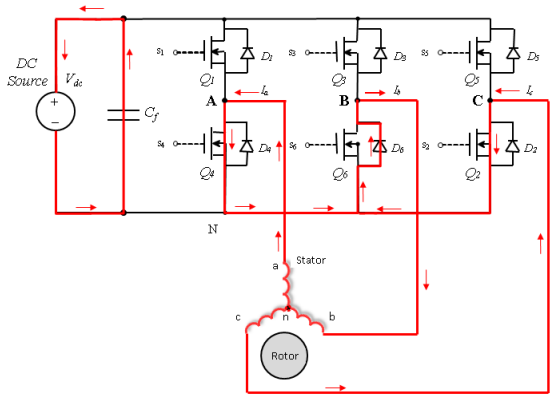
Therefore,

$$i_d = 0. \quad (10)$$

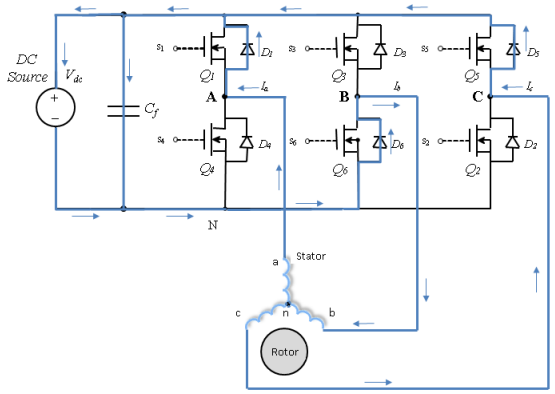
Substituting this result in equation (7) gives



(a)



(b)



(c)

Figure 3: Regenerative braking scheme. (a) Emf and PWM switching signals, (b) Current flow during ON-Time and 0-30° period. (c) Current flow during OFF-Time and 0-30° period

$$i_q = -\frac{3P\lambda_{pm}}{4R}\omega. \tag{11}$$

The equivalent electromagnetic torque is then given by

$$T_{em} = -\frac{\left(\frac{3}{4}p\lambda_{pm}\right)^2}{R}\omega. \tag{12}$$

This equation defines the minimum electromagnetic torque needed by the PMSM to operate the machine in the regenerative braking region.

Next, to find the maximum regenerative braking current absorbed, the input power is minimized.

$$\nabla P_{in} = \begin{bmatrix} \frac{\partial P_{in}}{\partial i_d} \\ \frac{\partial P_{in}}{\partial i_q} \end{bmatrix} = \begin{bmatrix} 2Ri_d \\ \frac{3}{2}\frac{p}{2}\lambda_{pm}\omega + 2Ri_q \end{bmatrix}. \tag{13}$$

The minimum power is obtained by setting this gradient to zero and solving for both variables, i_d and i_q.

$$i_d = 0, \tag{14}$$

$$i_q = -\frac{3p\lambda_{pm}}{8R}\omega. \tag{15}$$

The equivalent electromagnet torque generated by these current commands is given by

$$T_{em} = -\frac{9p^2\lambda_{pm}^2}{32R}\omega. \tag{16}$$

This torque guarantees maximum absorbed current by the DC source during regenerative braking.

3. Regenerative Braking with the Maximum Energy Recovery Switching Scheme (MERSS)

The switching scheme of the inverter is developed with the objective of maximizing energy recovery during regenerative braking. The idea is to utilize the motor phase inductors along with the inverter switches, functioning as a boost converter and allowing the phase currents to reverse their direction to flow back to the DC-bus capacitor [12], [14]. Regenerative braking is achieved by controlling only the lower switches (Q_2, Q_4, Q_6) through PWM and switching OFF all of the upper switches (Q_1, Q_3, Q_5) as shown in Fig. 3. In this mode of operation, the three lower switches are controlled with the same PWM command signal. During the ON-time, a path is provided for the phase current to flow in the negative direction through the closed switch, or in the positive direction through the free-wheeling diode depending on the polarity of the back emf.

Figure 3a shows the phase relationship between the back EMF, armature current, and the switching signals. Figure 3b and Figure 3c both show the closed loop path of the 3-phase currents during the 0-30s time interval of each cycle. During the OFF-time, the phase currents must maintain their direction and are therefore forced to flow through an alternate path created by the upper free-wheeling diodes, $D_1 - D_3$ and then back to the DC-bus capacitor. In this time period, regenerative braking is achieved and the capacitor is charged by the recovered electrical energy.

4. Experimental Testing and Discussions

In this section, experimental results of the regenerative braking process are analyzed to confirm the theoretical analysis. The energy recovered by the DC-bus capacitor is compared to the motor's mechanical energy to evaluate the efficiency of the

regenerative braking process. Figure 4 shows the topology of the PMSM experimental setup. The system consists of a PMSM, which is controlled using a MYWAY MWINV-9R144 inverter. The inverter switches are controlled using the dSPACE 1103 board. The PMSM is coupled with a Bühler DC Motor through flexible couplings and additional disc inertia mounted on the same shaft. The DC motor acts as a mechanical load and is controlled using a DC-DC converter, which in turn is controlled using a dSPACE 1104 board. Two encoders are used: the first is an incremental encoder directly connected to the DC motor side, and the second is a sine/cosine encoder connected to the PMSM side. The system parameters are listed in Table 1.

In the following sections, regenerative braking is implemented on the PMSM machine controlled with FOC. Two methods are implemented to realize regenerative braking. The first method is based on speed control and uses a ramp speed reference with negative slope to brake the motor. This linear deceleration simulates the braking of an EV, where the speed decreases gradually until it reaches 0. The braking time is adjusted by controlling the slope of the reference speed.

The second method operates the PMSM in torque control mode. The speed control loop is disconnected from the vector control scheme and the current commands, i_d^* and i_q^*, are generated directly. The i_q^* command will generate a negative torque that brings the motor to a stop. As discussed in section II-B, generating this torque guarantees maximum current absorption by the DC source, allowing for the returned current to achieve a higher amplitude and therefore the recovered energy is maximized.

Table 1: Motor drive system parameters

Parameter	*Value*
Resistance (R_{p-p})	6.8 Ω
Total inertia (J)	0.00315 Kg.m^2/s^2
Back EMF constant (K_b)	98 $V/krpm$
Torque constant (K_t)	1.6 Nm/A
Rated Torque	3.9 Nm
Stall Current	2.7 A
Inductance (L_{p-p})	24.3 mH
Damping coefficient (B)	4.741x10^{-4} Nm/(rad/s)
Coulomb friction (τ_c)	0.1343 Nm

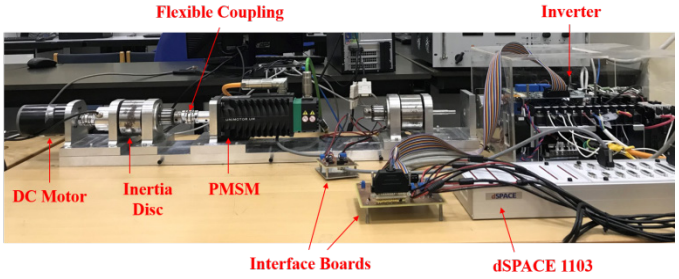
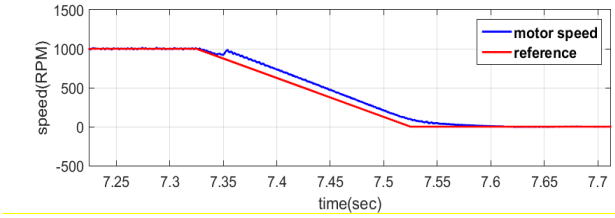


Figure 4: PMSM drive system

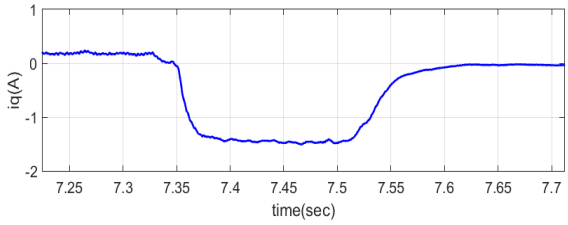
In order to observe the voltage increase during regenerative braking, the three-phase line is disconnected from the inverter and the DC-link voltage is maintained only by the DC-bus capacitor. As a result, the voltage begins decreasing. In this period of time, the motor is operating under constant speed. Once the DC-link voltage reaches 300V, the regenerative braking command signal is triggered and the motor operates under braking mode.

4.1. Regenerative braking using speed control mode

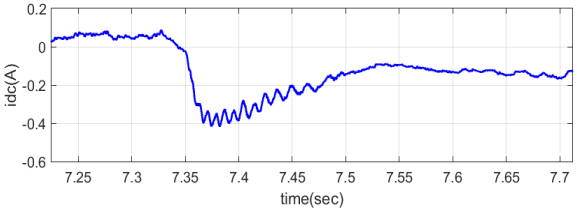
Figure 5a shows the reference speed and actual speed of the PMSM before and after regenerative braking is activated. The motor decelerates gradually to zero with a breaking time set to 0.2s. During the regenerative braking period, a negative q-axis current, i_q^*, is generated by the FOC controller as shown in Figure 5b. The motor mechanical energy is supplied back to the capacitor. Figure 5c shows a negative DC-link current indicating that energy is flowing from the motor to the capacitor. Figure 5d shows that the capacitor voltage is being charged during this period.



(a) Reference speed and actual speed



(b) q-axis current



(c) DC-link current

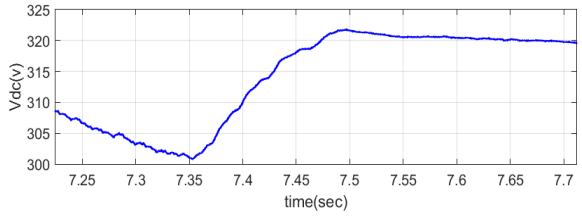


Figure 5: PMSM transient response during regenerative braking under the speed control mode of operation

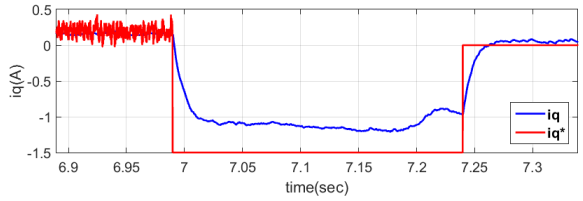
Table 2 gives a summary of the mode of operation for this form of regenerative braking. The total recovered energy is 70.496% of the motor energy.

Table 2: Summary of the regenerative braking performance for the speed control mode of operation

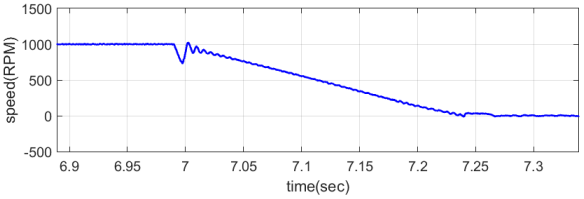
ω_0 (rpm)	Max i_{dc} (A)	ΔV_{dc} (V)	Recovery time (t_r)	Recovered Energy (J)	Braking Power (W)	Max i_q (A)	Mech. Energy (J)	Efficiency (%)
1000	−0.419	21.4	0.145	12.18	128.5	−1.508	17.27	70.50

Table 3: Summary of the regenerative braking statistics of the torque control mode of operation

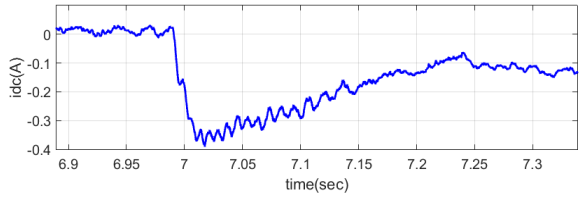
ω_0 (rpm)	Max i_{dc} (A)	ΔV_{dc} (V)	Recovery time (t_r)	Recovered Energy (J)	Braking Power (W)	Max i_q (A)	Mech. Energy (J)	Efficiency (%)
1000	−0.391	21.6	0.242	16.48	119.3	−1.218	17.27	95.43



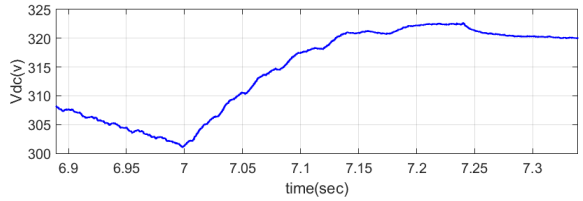
(a) Actual and command q-axis current



(b) Motor speed



(c) DC-link current



(d) DC-bus voltage

Figure 6: PMSM transient response during regenerative braking under the torque control mode of operation

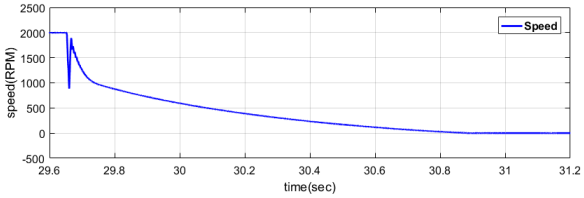
4.2. Regenerative braking using torque control

Figure 6 shows the motor variables during the torque mode of operation. Initially, the motor speed is regulated by FOC to the desired reference value. Once regenerative braking is initiated, the speed controller is disabled and the motor torque is controlled through the q-axis current. Figure 6a shows the reference current, i_q^*, and the generated q-axis current, i_q.

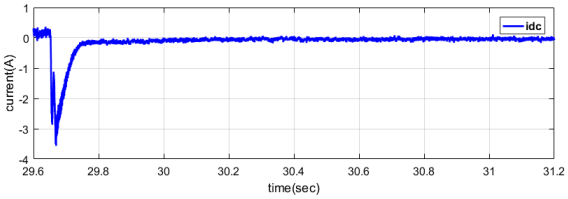
The braking time must be tuned in order for the speed to reach zero at the end of the breaking time. Table 3 summarizes the regenerative braking event results under the torque control mode of operation. The results indicate that greater amounts of energy are harvested with the torque control operational mode in comparison to the case of the speed controlled braking.

4.3. Regenerative braking using the Maximum Energy Regeneration switching scheme (MERSS)

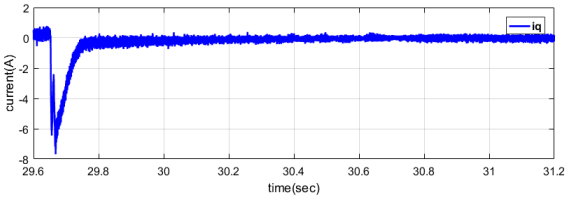
To evaluate the PMSM performance with MERSS, the motor is initially powered by the DC-bus capacitor to operate the drive system in driving mode. Next, a break command is activated to operate the system in regenerative braking mode.



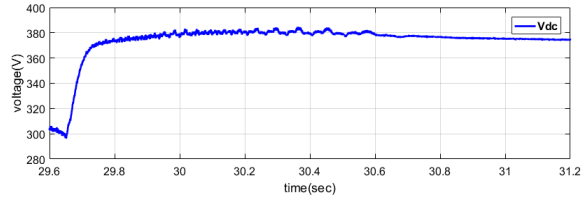
(a) Motor speed



(b) q-axis current



(c) DC-link current



(d) DC-bus voltage

Figure 7: PMSM transient response during regenerative braking under MERSS mode of operation. ω_0=2000 rpm, d=0.7

As a result, the phase currents reverse their direction and provide energy back to the DC-bus capacitor. The DC-bus voltage increases allowing the DC-bus capacitor to be charged.

Figure 7 shows the transient response of motor speed, q-axis motor current, DC-bus current, and DC-bus voltage. The motor is initially running at constant speed and the DC-bus capacitor is supplying power. When the break command is received, the motor starts decelerating until it comes to rest. The regenerative braking region is a subset of this period where the kinetic energy of the

rotor is used to generate electrical energy, and the motor acts as a generator. Figure 8 shows the electric power and energy recovered by the DC-bus capacitor. Following the braking command, energy recovery is activated through the MERSS and the capacitor voltage increases as current is absorbed by the capacitor. Regenerative breaking stops when the motor current reaches the minimum level set by the speed, as given through equation (11).

This process was repeated for different motor speeds and different PWM duty cycles. Varying the duty cycle of the brake command will change the ON-OFF times of the lower inverter switches, as explained previously in Figure 3. Increasing the ON-time period will charge the motor's inductance for a longer time allowing for the storage of additional energy. This process is similar to the operation of a boost converter. However, if it the ON-period is heavily increased, energy will not be fully recovered due to a short OFF-time period, and will eventually be lost in the switching process. For this reason, the duty cycle must be carefully selected for optimum energy recovery. Figure 9 shows the regenerated power and energy as a function of duty cycle for different motor speeds.

The maximum regenerated energy is compared with the mechanical energy to evaluate the efficiency of the system under MERSS. The energy of the motor and the energy stored in the ultracapacitor during braking are given by:

$$\mathrm{E_{mech}} = \frac{1}{2}\mathrm{J}\omega_0^2, \tag{17}$$

$$E_{cap}(k) = E_{cap}(k-1) + \frac{T_s}{2}[P_{cap}(k) + P_{cap}(k-1)], \tag{18}$$

where, ω_0 is the initial speed of the motor when the braking process starts. Therefore, the efficiency of the braking process takes into consideration the power losses during switching.

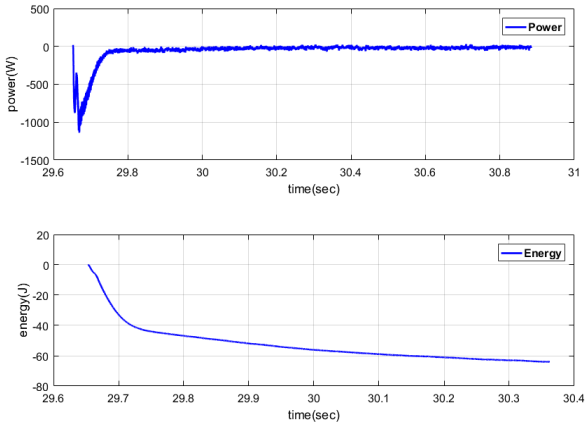


Figure 8: Electric power and energy recovered by the DC bus with MERSS. ω_0=2000 rpm, d=0.7

Table 4 summarizes the results from this experiment and displays the efficiency, as well as the energy harvested. It can be observed that the optimum duty cycle that yields maximum energy recovery is a function of the speed.

The performance of the MERSS is next compared to FOC by conducting the regenerative braking experiments with FOC under the same operating conditions. Table 5 shows a summary of the results. It can be observed that MERSS yields always a better efficiency than FOC.

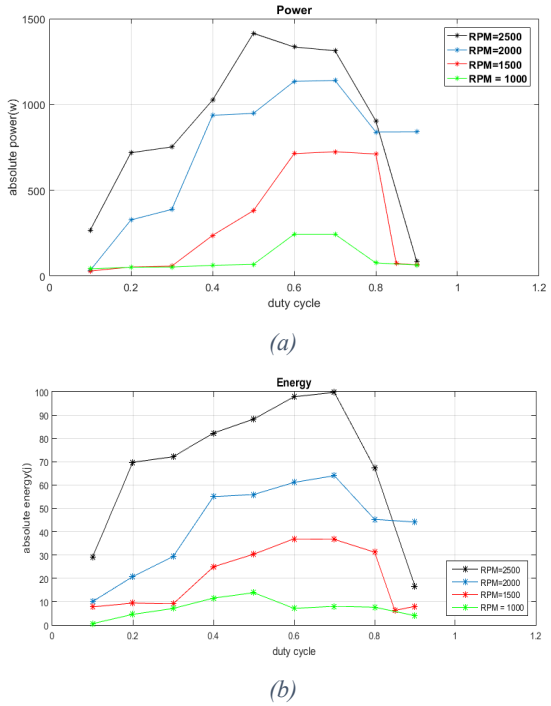


Figure 9: Electric power and energy recovered on the DC bus with MERSS

$$\eta = \frac{E_{cap}}{E_{mech}}. \tag{19}$$

Table 4: Maximum efficiency of braking at different motor speeds with MERSS.

ω_0 (rpm)	Duty Cycle	$E_{mech}(J)$	E_{cap} (J)	η (%)
1000	0.5	17.27	13.84	80.14
1500	0.6	38.86	36.96	95.11
2000	0.7	69.09	64.00	92.64
2500	0.7	107.95	99.74	92.40

Table 5: Comparative analysis between FOC and MERSS

	ω_0 (rpm)	Max i_{dc} (A)	ΔV_{dc} (V)	Max i_q (A)	E_{mech} (J)	E_{cap} (J)	Efficiency (%)
FOC	1000	-0.419	21.4	-1.508	17.27	12.18	70.50
	1500	-1.005	54.9	-2.312	38.86	27.73	71.36
	2000	-1.575	96.8	-2.958	69.09	50.62	73.27
MERSS	1000	-0.2203	16.6	-0.740	17.27	13.84	80.14
	1500	-2.13	52.0	-4.968	38.86	36.96	95.11
	2000	-3.564	85.9	-7.683	69.09	64.00	92.64

5. Conclusion

This paper discusses regenerative braking in a PMSM drive system. A new maximum energy recovery switching scheme is developed and compared to FOC. Analysis is carried out on an experimental setup to confirm the effectiveness of regenerative braking under the new scheme. Experimental results show that the maximum current absorbed by the ultracapacitor does not exceed a set limit which is in turn dependent on the ultracapacitor's voltage, internal resistances of the system and the motor speed. The effect of varying the duty cycle is also studied to uncover the

highest efficiency duty cycle for optimum harvesting of regenerated energy.

References

[1] M. K. Yoong, Y. H. Gang, G. D. Gan, C. K. Leong, Z. Y. Phuan, B. K. Cheah, and K. W. Chew, "Studies of regenerative braking in electric vehicle" in 2010 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT), Petaling Jaya, Malaysia, 2010. https://doi.org/10.1109/STUDENT.2010.5686984.

[2] M.-Ji Yang, H.-L. Jhou, B.-Y. Ma, and K.-K. Shyu, "A Cost-Effective Method of Electric Brake With Energy Regeneration for Electric Vehicles" Ming-Ji Yang, Hong-Lin Jhou, Bin-Yen Ma, and Kuo-Kai Shyu, Member, IEEE Trans. Ind. Electronics, 56(6), 2203 - 2212, June 2009. https://doi.org/10.1109/TIE.2009.2015356.

[3] D. Lu, M. Ouyang, J. Gu, and J. Li, "Instantaneous optimal regenerative braking control for a permanent-magnet synchronous motor in a four-wheel-drive electric vehicle" Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 228(8), 894-908, July 2014. https://doi.org/10.1177/0954407014521173.

[4] C.-H. Chen, W.-C. Chi, and M.-Y. Cheng, "Regenerative braking control for light electric vehicles" in IEEE 9th International Conference on Power Electronics and Drive Systems (PEDS), Singapore, 2011. http:// https://doi.org/10.1109/PEDS.2011.6147317.

[5] F. Naseri; E. Farjah; T. Ghanbari, "An Efficient Regenerative Braking System Based on Battery/Ultracapacitor for Electric, Hybrid and Plug-In Hybrid Electric Vehicles with BLDC Motor" IEEE Trans. Vehicular Technology, 66(5), 3724–3738, 2017. https://doi.org/10.1109/TVT.2016.2611655.

[6] Y. Bian, L. Zhu. H. Lan, A, Li. Anhu, and X. Xu, "Regenerative Braking Strategy for Motor Hoist by Ultracapacitor" Chin. J. Mech. Eng., 25(2), 377–384, March 2012. https://doi.org/10.3901/CJME.2012.02.377.

[7] Z. Zhang, X. Zhang, W. Chen, Y. Rasim, W. Salman, H. Pan, Y. Yuan, and C. Wang, "A high-efficiency energy regenerative shock absorber using supercapacitors for renewable energy applications in range extended electric vehicle" Applied Energy, 178, 177–188, 2016. https://doi.org/10.1016/j.apenergy.2016.06.054.

[8] S. Ding, M. Cheng, Chao Hu, Guishu Zhao and Wei Wang, "An energy recovery system of regenerative braking based permanent magnet synchronous motor for electric vehicles" in 2013 International Conference on Electrical Machines and Systems (ICEMS), Busan, South Korea, 2013. https://doi.org/10.1109/ICEMS.2013.6754468.

[9] Ned Mohan, Advanced Electric Drives: Analysis, Control and Modeling Using Simulink, Wiley 2014.

[10] A. Samba Murthy, "Analysis of regenerative braking in electric machines," MSc Thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, May 2013.

[11] A. Adib and R. Dhaouadi, "Modeling and analysis of a regenerative braking system with a battery-supercapacitor energy storage" in 7th International Conference on Modeling, Simulation, and Applied Optimization (ICMSAO), Sharjah, UAE, 2017. https://doi.org/10.1109/ICMSAO.2017.7934897.

[12] A. S. Murthy, D. P. Magee and D. G. Taylor, "Vehicle braking strategies based on regenerative braking boundaries of electric machines" in 2015 IEEE Transportation Electrification Conference and Expo (ITEC), Dearborn, MI, 2015. https://doi.org/10.1109/ITEC.2015.7165809.

[13] T.-H. Kim, J.-H. Lee, and C.-Y. Won, "Design and control methods of bidirectional DC-DC converter for the optimal DC link voltage of PMSM drive" Journal of Electrical Engineering and Technology, 9(6), 1944-1953, 2014. http://dx.doi.org/10.5370/JEET.2014.9.6.1944.

[14] X. Jiaqun and C. Haotian, "Regenerative brake of brushless DC motor for light electric vehicle" in 2015 18th International Conference on Electrical Machines and Systems (ICEMS), Pattaya, Thailand, 2015. https://doi.org/10.1109/ICEMS.2015.7385262.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 467-471 (2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Tracking and Detecting moving weak Targets

Naima Amrouche[*,1,2], Ali Khenchaf[2], Daoud Berkani[1]

[1]*Ecole Nationale Polytechnique d'Alger, 10 Rue Frère Oudek, Elharrach, Algérie, 16200*

[2]*Lab-STICC CNRS UMR 6285, ENSTA Bretagne, 02 Rue François Verny, Brest Cedex 09, France, 29806*

ARTICLE INFO

Article history
Received: 13 December, 2017
Accepted: 24 January, 2018
Online: 10 February, 2018

Keywords :
Target Tracking
Track Before Detect
Detection
Particle Filter

ABSTRACT

Detect and tracking of moving weak targets is a complicated dynamic state estimation problem whose difficulty is increased in case of high clutter conditions or low signal to noise ratio (SNR). In this case, the track-before-detect filter (TBDF) that uses unthresholded measurements considers as an effective method for detecting and tracking a single target under low SNR conditions. In this paper, a particle filter based track-before-detect (PF-TBD) method is proposed to address the problem of detection and tracking with unthersholded data and a binary variable of the existence of the target for two motion models. Simulation results using image measurements based on TBD scenarios are also presented to demonstrate the capability of the proposed approach.

1. Introduction

The classical approach to target tracking is based on target measurements (position, range rate, and so forth) that are extracted by thresholding the output of a signal processing unit of a surveillance sensor [1].The primary role of thresholding is to reduce the data flow and thus simplify tracking. For a target of a certain signal-to-noise ratio (SNR), the choice of the detection threshold determines the probability of target detection and the density of false alarms. The false alarm rate, on the other hand, affects the complexity of the data association problem in the tracking system. In general, higher densities of false alarms require more sophisticated data association algorithms.

The undesirable effect of thresholding the sensor data, however, is that in restricting the data flow, it also throws away potentially useful information. For high SNR targets this loss of information is of little concern because one can achieve good probability of detection with a small false alarm rate. Recent developments of stealthy military aircraft and cruise missiles have emphasized the need to detect and track low SNR targets. For these dim (stealthy) targets, there is a considerable advantage in using the unthresholded data for simultaneous detection and track initiation [2], [3]. Depending on the type of sensor in use, the unthresholded data can be a sequence of range-Doppler maps, bearing-frequency distributions.

[*]Naima Amrouche, 02 rue François Verny, Tel : +33 2 98 34 88 00 &
Email: naima.amrouche@ensta-bretagne.org

The concept of simultaneous detection and tracking using unthresholded data is known in literatures as track-before-detect (TBD) approach. Typically TBD is implemented as a batch algorithm using the Hough transform [4], dynamic programming [2] [3] or maximum likelihood estimation [5].

TBD algorithms based on the Hough transformation, dynamic programming or maximum likelihood methods are generally computationally intensive [6]. With recent advancement in Sequential Monte Carlo techniques, TBD algorithms implemented using PF are now computationally feasible [7] [8].

In this paper we also develop a recursive Bayesian TBD estimator; however, our formulation and implementation are based on the particle filter [7] [9]. The PF based TBD incorporates unthresholded data and a binary target existence variable into the target state estimation process. The presence and absence of target are explicitly modelled [10] [11]. This concept, allows us to calculate the probability of existence of the target directly from the filter.

The paper is organized as follows: section 2 the system dynamics and measurement model, are introduced for the TBD application. In section 3 formulates the TBD approach as a nonlinear filtering problem and describes the conceptual recursive Bayesian solution. The implementation of this solution using a particle filter is presented in section 4. In section 5 collects our simulations and results. Finally, we report our conclusions and direction for future research in section 6.

2. Formulation Problem

2.1. Dynamic Model

We assume that want to track a target moving in a 2-D plane with an unknown state vector s_k at time step k. We consider the state model given by:

$$s_{k+1} = Fs_k + v_k \quad (1)$$

Where F is the state transition matrix, assuming constant velocity motion and coordinate turn motion respectively, k is the discrete-time index, v_k is the process noise sequence, and s_k is the state vector defined as:

$$s_k = [x_k \quad \dot{x}_k \quad y_k \quad \dot{y}_k \quad I_k] \quad (2)$$

Here $(x_k, y_k), (\dot{x}_k, \dot{y}_k)$ and I_k denote the position, velocity, and the intensity of the target, respectively.

2.2. Transition Matrix

A target can be present or absent from the surveillance region at a discrete-time k. Target presence variable E_k is modelled by a two-state Markov chain, that is $E_k = \{0,1\}$. Here 0 denotes the event that a target is not present, while 1 denotes the opposite. Furthermore, we assume that transitional probabilities of target "birth" P_b and "death" P_d, defined as:

$$P_b \triangleq P\{E_k = 1 | E_{k-1} = 0\} \quad (3)$$

$$P_d \triangleq P\{E_k = 0 | E_{k-1} = 1\} \quad (4)$$

Are known, the other two transitional probabilities of this Markov chain, the probability of staying alive $1 - P_d$ and the probability of remaining absent $1 - P_b$ respectively, are given by:

$$1 - P_d \triangleq P\{E_k = 1 | E_{k-1} = 1\} \quad (5)$$

$$1 - P_b \triangleq P\{E_k = 0 | E_{k-1} = 0\} \quad (6)$$

The corresponding transition matrix for the Markov process is:

$$\Pi = \begin{bmatrix} 1 - P_b & P_b \\ P_d & 1 - P_d \end{bmatrix} \quad (7)$$

2.3. Sensor Model

The sensor provides a sequence of two-dimensional images (frames) of the surveillance region, each image consisting of $(n \times m)$ resolution cells (pixels). A resolution cell corresponds to a rectangular region of dimensions $\triangle_x \times \Delta_y$ so that the center of each cell (i, j) is defined to be at $(i\Delta_x \times j\Delta_y)$ for $i = 1, \dots n$ and $= 1, \dots m$.

At each resolution cell (i, j) the measured intensity is denoted as $z_k^{(i,j)}$ and modeled as:

$$z_k^{(i,j)} = \begin{cases} h_k^{(i,j)}(s_k) + \omega_k^{(i,j)} & if\ target\ present \\ \omega_k^{(i,j)} & if\ target\ absent \end{cases} \quad (8)$$

Where $h_k^{(i,j)}(s_k)$ is the target contribution to intensity level in the resolution cell (i, j) and $\omega_k^{(i,j)}$ is measurement noise in the resolution cell (i, j), assumed to be independent from pixel to pixel and from frame to frame. Thus for a point target of intensity I_k at position(x_k, y_k), the contribution to pixel(i, j) is approximated as:

$$h_k^{(i,j)}(s_k) \approx \frac{\Delta_x \Delta_y I_k}{2\pi\Sigma^2} \exp\left\{-\frac{(i\Delta_x - x_k)^2 + (j\Delta_y - y_k)^2}{2\Sigma^2}\right\} \quad (9)$$

Where Σ is the amount of blurring introduced by the sensor. The complete measurements recorded at time k a $n \times m$ matrix denoted as:

$$z_k = \left\{z_k^{(i,j)} : i = 1, \dots, n, j = 1, \dots, m\right\} \quad (10)$$

While the set of complete measurements collected up to time k is denoted as usual: $Z_k = \{z_i, i = 1, \dots k\}$.

3. Bayesian Solution to TBD Filtering

The formal recursive Bayesian solution can be presented as a two-step procedure, consisting of prediction and update.

3.1. Prediction

The predicted target state can be written in terms of the target state and existence at the previous time, giving

$$\begin{gathered} p(s_k, E_k = 1 | Z_{k-1}) = (1 - \\ P_d) \int p(s_k | s_{k-1}, E_k = 1, E_{k-1} = 1) \times \\ p(s_{k-1}, E_{k-1} = 1 | Z_{k-1}) ds_{k-1} + \\ P_b \int p_b(s_k) p(s_{k-1}, E_{k-1} = 0 | Z_{k-1}) ds_{k-1} \end{gathered} \quad (11)$$

The pdf $p_b(s_k)$ denotes the initial target density on its appearance.

3.2. Update

The update equation in the Bayesian framework is given by:

$$\begin{gathered} p(s_k, E_k = 1 | Z_k) = \\ \frac{p(z_k | s_k, E_k = 1) p(s_k, E_k = 1 | Z_{k-1})}{p(z_k | Z_{k-1})} \end{gathered} \quad (12)$$

Where prediction density $p(s_k, E_k = 1 | Z_{k-1})$ is given by (11) and $p(z_k | s_k, E_k)$ is the likelihood function given by:

$$\begin{gathered} p(z_k | s_k, E_k) = \\ \begin{cases} \prod_{i=1}^{n} \prod_{j=1}^{m} p_{S+N}\left(z_k^{(i,j)} \middle| s_k\right), & for \quad E_k = 1 \\ \prod_{i=1}^{n} \prod_{j=1}^{m} p_N\left(z_k^{(i,j)}\right), & for \quad E_k = 0 \end{cases} \end{gathered} \quad (13)$$

Here $p_N\left(z_k^{(i,j)}\right)$ is the probability density function of background noise in pixel (i, j), while $p_{S+N}\left(z_k^{(i,j)} \middle| s_k\right)$ is the likelihood of target signal plus noise in pixel (i, j),given that the target is in state s_k, This two probability density function can be further expressed as:

$$p_N\left(z_k^{(i,j)}\right) = \mathcal{N}\left(z_k^{(i,j)}, 0, \sigma^2\right) \quad (14)$$

$$p_{S+N}\left(z_k^{(i,j)}\middle|s_k\right) = \mathcal{N}\left(z_k^{(i,j)}, h_k^{(i,j)}, \sigma^2\right) \quad (15)$$

Since the target (if present) will affect only the pixels in the vicinity of its location (x_k, y_k), the expression for $p(z_k|s_k, E_k = 1)$ can be approximated as follows:

$$p(z_k|s_k, E_k = 1) \approx \prod_{i\in C_i(s_k)} \prod_{j\in C_j(s_k)} p_{S+N}\left(z_k^{(i,j)}\middle|s_k\right) \cdot \prod_{i\notin C_i(s_k)} \prod_{j\notin C_j(s_k)} p_N\left(z_k^{(i,j)}\right) \quad (16)$$

Where $C_i(s_k)$ and $C_j(s_k)$ are the sets of subscripts i and j, respectively, corresponding to pixels affected by the target.

4. A Particle Filter for Track Before Detect (PF-TBD)

The recursive Bayesian solution of the track problem described in the previous section can be implemented using a particle filter [7] [9] [12] [13]has some similarities to the MMPF [14]. In this case we introduce the augmented state vector to include the existence variable. $y_k = [s_k^T \quad E_k]^T$. Let us denote a random measure that characterizes the posterior probability density function at $k-1$, namely $p(y_{k-1}|Z_{k-1})$, by $\{y_{k-1}^n, \omega_{k-1}^n\}_{n=1}^N$.As usual, N is the number of particles, while y_k^n consists of s_k^n and E_k^n.The pseudocode of a single cycle of the PF developed for the TBD problem is presented in Table 1. The next step is the prediction of particle target states; this is done, however, only for those particles that are characterized by $E_k^n = 1$.For remaining particles (with $E_k^n = 0$), the target state components are undefined. There are two possible cases here:

4.1. Newborne Particles

This group of predicted particles is characterized by the transition from $E_{k-1}^n = 0$ to $E_k^n = 1$.The target state particles are uniformly drawn at time step k based on some a priori information on the minimum and maximum possible values on the target state.

4.2. Existing Particles

This group of particles that continues to stay "alive", with $E_{k-1}^n = 1$ to $E_k^n = 1$.The state transition model in equation (1) is used to update the target state particles.

The importance weights are computed next. For this purpose we need to introduce the likelihood ratio in pixel (i,j) for a target in state s_k^n, defined as:

$$\ell\left(z_k^{(i,j)}\middle|s_k^n\right) \triangleq \frac{p_{S+N}\left(z_k^{(i,j)}\middle|s_k^n\right)}{p_n\left(z_k^{(i,j)}\right)} \quad (17)$$

$$\ell\left(z_k^{(i,j)}\middle|s_k^n\right) = \exp\left\{-\frac{h_k^{(i,j)}\left(h_k^{(i,j)} - 2z_k^{(i,j)}\right)}{2\sigma^2}\right\} \quad (18)$$

Where $h_k^{(i,j)}$ was defined in (9). Equation (18) follows from (14), (15), and (11). The importance weights (up normalizing constant) are now given by [7]:

$$\tilde{\omega}_k^n = \begin{cases} \prod_{i\in C_i(s_k^n)} \prod_{j\in C_j(s_k^n)} \ell\left(z_k^{(i,j)}\middle|s_k^n\right) & if \quad E_k^n = 1 \\ 1 & if \quad E_k^n = 0 \end{cases} \quad (19)$$

$[\{y_k^n\}_{n=1}^N]$ =TBD-PF$\left[\{y_{k-1}^n\}_{i=1}^N, z_k\right]$

- Target existence transitions using the Regime Transition Algorithm given in [6]

$[\{E_k^n\}_{n=1}^N]$ =RT$[\{E_{k-1}^n\}_{n=1}^N, \Pi]$

- FOR $i = 1{:}\,N$
  - IF a newborn particle ($E_{k-1}^n = 0$ and $E_k^n = 1$) Draw $s_k^n \sim q_b(s_k|z_k)$
  - IF an existing particle ($E_{k-1}^n = 1$ and $E_k^n = 1$) Draw $s_k^n \sim q(s_k|s_{k-1}^n, z_k)$
  - Evaluate importance weight using (13)
- END FOR
- Calculate total weight: t =SUM$[\{\tilde{\omega}_k^n\}_{n=1}^N]$
- FOR $n = 1{:}\,N$
  - Normalize: $\omega_k^n = t^{-1}\tilde{\omega}_k^n$
- END FOR
- Resample using systematic resampling algorithm given in [6]

$[\{y_k^n, -, -\}_{n=1}^N]$ =RESAMPLE$[\{y_k^n, \omega_k^n\}_{n=1}^N]$

The PF for track-before-detect performs target detection using the estimate of the posterior probability of target existence. This estimate is computed as:

$$\hat{P}_k = \frac{\sum_{n=1}^N E_k^n}{N} \quad (20)$$

And satisfies $0 < \hat{P}_k < 1$. Target presence is declared if $\hat{P}_k$ is above a certain threshold value. This declaration can then trigger the initialization of a track based on the estimated target state

$$\hat{s}_{k/k} = \frac{\sum_{n=1}^N s_k^n \cdot E_k^n}{\sum_{n=1}^N E_k^n} \quad (21)$$

5. Simulation and Results

In the simulation we used two scenarios for target motion and random walk model is adopted for target intensity.

5.1. Scenario1

The first model is a nearly constant velocity model is used. The dynamic model [15] for the target can be described by (1).

Where $F = diag[F_1, F_1]$,$Q = diag[Q_1, Q_1, q_2 T]$

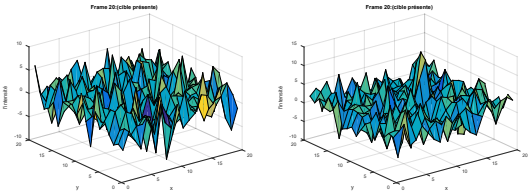
and $F_1 = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$, $Q_1 = q_1 \cdot \begin{bmatrix} T^3/3 & T^2/2 \\ T^2/2 & T \end{bmatrix}$

Where q_1 and q_2 denote the level of process noise in target motion and intensity, respectively. A sequence of 30 frames of data has been generated with the following parameters:

$\Delta_x = \Delta_y = 1, n = m = 20, T = 1s, \sigma = 3, \Sigma = 0.7$. The target is absent from frame 1 to frame 5 to be introduced in frame 6 with the initial intensity I_0 .The initial state is $[4.2 \quad 0.45 \quad 7.2 \quad 0.25 \quad I_0]$ [6] [16] [17].

The simulations are conducted under an initial intensity $I_0 = 9, 13$ and 25, which corresponds to an SNR of 3.18,6.71, and 12 dB, respectively, according to the calculation equation $SNR = 10log\left[\frac{I\Delta_x\Delta_y/2\pi\Sigma^2}{\sigma}\right]^2$. The target exists until frame 24 and is again absent in frames 25, 26,…, 30.

Figure 1 (a) and (b) show the measurement frame at time step 20 for 6.71dB and 12 dB peak respectively.



(a) (b)

Fig. 1. Measurements Frame (20): (a) for 6.71 dB Peak SNR, (b) for 12 dB Peak SNR for CV model.

The particle filter parameters are selected as follows: transitional probabilities $P_b = P_d = 0.05$;initial existence probability $\mu_1 = 0.05; v_{\min} = -1\ unit/s$; $v_{\max} = 1\ unit/s$; initial intensity range from $I_{min} = I_0 - 5$ to $I_{max} = I_0 + 5$; $p = 2$ and number of particles $N = 2000$.

In figure (2) the probability of presence is shown for a SNR of 6.71 dB. Existence probability remains very stable and above 0.97 until frame 25.Then it drop sharply in frame 26, when the target is disappear.

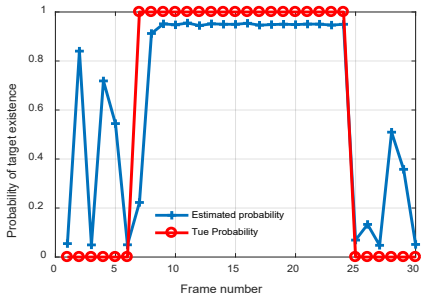


Fig. 2. True and Estimated Target Existence Probability for SNR=6.71 dB

Figure (3) displays the true target path against the track, produced by the filter. Note how the target trajectory deviates slightly from the straight line due to process noise. The PF-TBD tracks the target with a small positional error.

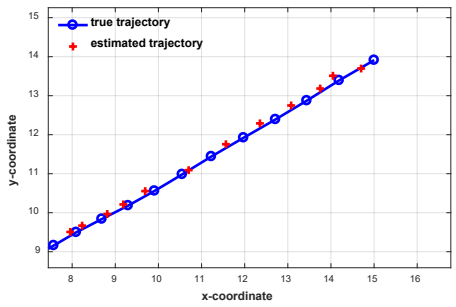


Fig. 3. True and Estimated Target Trajectory for SNR=6.71 dB

Figure (4) shows the position RMSE for three different peak SNR conditions (3.18 dB,6.71 dB, 12 dB). The position error is lower in 6.71 dB than 3.18 dB. As it can be seen, the PF-TBD was able to closely track the target even under low SNR.

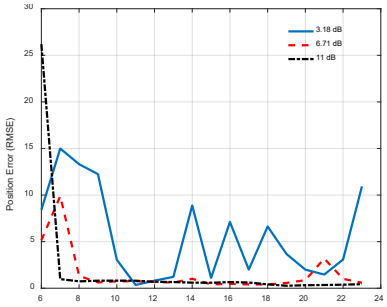


Fig. 4. Position RMSE for the PF-TBD for differnet peak SNR

5.2. Scenario2

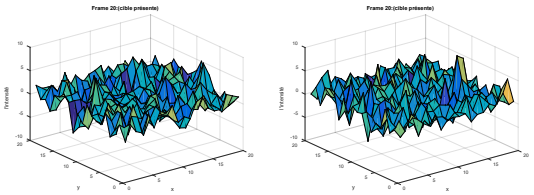
The second model is a Coordinate turn model is used [15]. The dynamic model for the target can be described by (1).

Where:

$$F = \begin{bmatrix} 1 & F_1 & 0 & F_2 & 0 \\ 0 & F_3 & 0 & -F_4 & 0 \\ 0 & -F_2 & 1 & F_1 & 0 \\ 0 & F_4 & 0 & F_3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, Q = \begin{bmatrix} Q_1 & Q_2 & 0 & Q_3 & 0 \\ Q_2 & Q_4 & -Q_3 & 0 & 0 \\ 0 & -Q_3 & Q_1 & Q_2 & 0 \\ Q_3 & 0 & Q_2 & Q_4 & 0 \\ 0 & 0 & 0 & 0 & Q_5 \end{bmatrix}$$

And $F_1 = \frac{sin(\Psi T)}{\Psi}$, $F_2 = \frac{(-cos(\Psi T)+1)}{\Psi}$, $F_3 = \cos(\Psi T)$, $F_4 = \sin(\Psi T)$, $Q_1 = \frac{2(\Psi T - sin(\Psi T))q_1}{\Psi^3}$, $Q_2 = \frac{(1-cos(\Psi T)q_1)}{\Psi^2}$ $Q_3 = \frac{(\Psi T - sin(\Psi T)q_1)}{\Psi^2}$, $Q_4 = q_1 T$, $Q_5 = q_2 T$, $\Psi = 6$ is a constant angular rate.

Figure 5 (a) and (b) show the measurement frame at time step 20 for 6.71dB and 12 dB peak respectively.



(a) (b)

Fig. 5. Measurements Frame (20): (a) for 6.71dB Peak SNR, (b) for 12dB Peak SNR for CT model.

In figure (6) the probability of presence is shown for a SNR of 6.71 dB. Existence probability probability is still increase above frame 7 until frame 17 and still stable until frame 25. Therefore, it drops rapidly following the target disappears from the monitoring region after frame 25.

Figure (7) shows the true and estimated target trajectories for coordinate turn model, the estimated trajectory is very close to the true trajectory.

Figure (8) shows the position RMSE for three different peak SNR conditions (3.18 dB,6.71 dB, 12 dB). The position error is lower in 6.71 dB than 3.18 dB. As it can be seen, the PF-TBD was able to closely track the target even under low SNR.

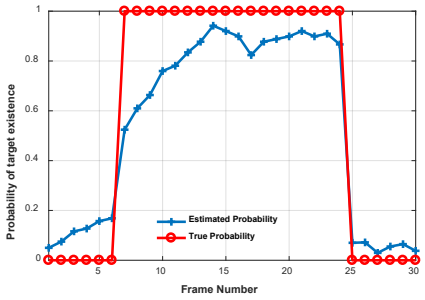


Fig. 6. True and Estimated Target Trajectory for SNR=6.71 dB

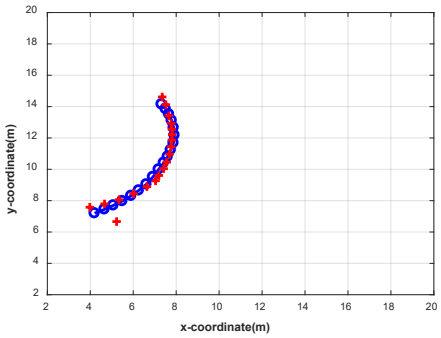


Fig. 7. True and Estimated Target CT Trajectory for SNR=6.71

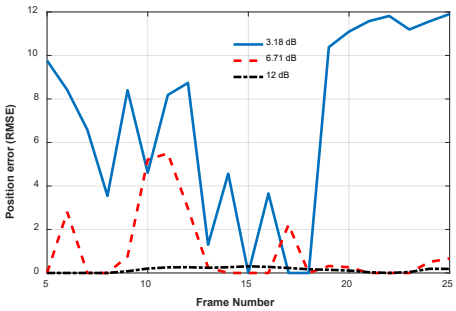


Fig. 8. Position RMSE for the PF-TBD for differnet peak SNR

6. Conclusion

In this paper, to manipulate moving weak targets, the PF-TBD algorithm is proposed for two dynamics models (CV and CT). The major advantage of the track-before-detect approach based on target existence variable and as a result, the developed particle filter can detect and track low SNR maneuvering target. The results from the simulation show that the PF-TBD algorithm has a successfully detection and tracking performance, both for constant velocity and coordinate turn models of moving targets, under severe conditions such as high noise or low SNR. Therefore, further work will mainly concentrate on how to detect and track multiple targets in high noise and high clutter.

References

[1] S. Blackman and R. Popoli, Design and Analysis of Modern Tracking System, Norwood: MA: Artech House, 1999.

[2] Y. Barniv, Dynamic programming algorithm for detecting dim moving targets, in Multitarget Multisensor Tracking: Advanced Application (Y, Bar Shalom, ed),ch 4, Norwood: MA, Artech House, 1990.

[3] J. Arnold, S. Shaw and H. Pasternack, "Efficient target tracking using dynamic programming," *IEEE Trans Aerospace and Electronic Systems,* vol. 29, pp. 44-56, January 1993.

[4] B. Carlson, E. D. Evans and S. L. Wilson, "Search radar detection and track with the Hough transform, part I: System concept," *IEEE Trans Aerospace and Electronic System,* vol. 30, pp. 102-108.

[5] M. Tonissen and Y. Bar-Shalom, "Maximum likelihood track-before-detect with fluctuating target amplitude," *IEEE Trans, Aerospace and Electronic Systems,* vol. 34, pp. 796-809, July 1998.

[6] R. Ristic, S. Arulampalam and N. Gordon, Beyond the Kalman Filter: Particle Filters for Tracking Applications, Boston: MA:Artech House, 2004.

[7] D. Salmond and H. Birth, "A particle filter for track-before-detect," *in Proc, Americain Control Conf,* pp. 3755-3760, June 2001.

[8] Y. Boyers and H. Drissen, "Particle filter based detection for tracking," *In Proceedings of the American Control Conference,,* pp. 4393-4397, June 2001.

[9] M. Rollason and D. Salmond, "A particle filter for track-before-detect of a target with unknown amplitude," *in IEE Int, Seminar Target Tracking: Algorithms and Application,* p. 14, October 2001.

[10] D. B. Colegrove, A. W. Davis and J. K. Alyliffe, "Track initiation and nearest neihbours incorporated into probabilistic data association," *Journal of Electrical and Engineers,* vol. 6, pp. 191-198, September 1986.

[11] D. Musicki, R. Evans and S. Stankovic, "Integrated probabilistic data association," *IEEE Trans. Automatic Control,* vol. 39, pp. 1237-1240, June 1994.

[12] D. J. Ballantyne, H. Y. Chan and M. A. Kouritzin, "A Novel branching particle method for tracking," *in Proc, SPIE, Signal and Data Processing of Small Targets,,* vol. 4048, p. 287, 2000.

[13] M. G. Rutten, B. Ristic and N. J. Gordon, "A Comparison of Particle Filters for Recursive Track-before-detect," in *7th International Conferenceon Information Fusion (FUSION)*, 2005.

[14] S. Mcginnity and G. W. Irwin, "Multiple Bootstrap Filter for Maneuvering Target Tracking," *IEEE Transaction of Aerospace and Electronic systems,* vol. 36, no. 3, July 2000.

[15] Y. Bar-Shalom, X. R. Li and T. Kirubarajan, Estimation with Applications to Tracking and Navigation, New York: Jhone Wiley&Sons, 2001.

[16] E. S. P and -S. A. P, "Generalized Recursive Track-Before-Detect With Proposal Partitioning for Tracking Varying Number of Multiple Targets in Low SNR," *IEEE Transactions on signal processing,* vol. 64, no. 11, 2016.

[17] E. P. S and -S. P. A, "Generalized Recursive Track-Before-Detect With Proposal Partitioning for Tracking Varying Number of Multiple Targets in Low SNR," *IEEE Transactions on signal processing,* vol. 64, no. 11, 2016.

[18] N. Amrouche, A. Khenchaf and D. Berkani, "Detection and Tracking Targets under Low SNR," in *IEEE International Conference on Industrial Technology*, Toronto, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 472-482 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Structure-Preserving Modeling of Safety-Critical Combinational Circuits

Feim Ridvan Rasim*, Sebastian M. Sattler

Chair of Reliable Circuits and Systems, Friedrich-Alexander-University Erlangen-Nuremberg, Paul-Gordan-Str. 5, 91052 Erlangen, Germany

ARTICLEINFO

Article history:
Received: 31 October, 2017
Accepted: 01 February, 2018
Online: 18 February, 2018

Keywords:
Combinational circuit
Structure
modeling
Symbolic analysis
Signal flow graph
Signal flow plan
Boolean function
Resolution method
KV diagram

ABSTRACT

In this work, a representative combinational circuit is visualized in various ways. It is abstracted (concretized) from transistor level to gate level and a structure-preserving transition is carried out into a signal flow graph. For creating a signal flow plan it is necessary to swap the nodes and the edges in the signal flow graph. After having executed this action the result is a signal flow plan. A value table exhibits the coding of the whole circuit. Then the so called module view is used to get the familiar compact and directed display and neighborhood relations are repeated once more, the resolution method is used. It is observed that undefined results can occur in digital circuits. But, these must be avoided in safety critical circuits. These events have to be secured in practice by costly and expensive verification and testing. In order to deal with the problem now, the structure-preserving modeling has to be understood, since this is the only way to achieve a one-purpose, qualitative and cost effective search for errors.

1 Introduction

This paper is an extension of work orginally presented at the 20th IEEE International Symposium on DDECS 2017, Dresden, Germany [1].

In order to ensure the functional safety of circuits or systems which are regarded as critical to safety, the mutual convert of models and functions is of great importance. The inconsistency problem is omnipresent; therefore, the essential claim for conformity with the formal derived function and the function derived from the real structure has a present role [2]. The directed mode of operation of a system should be represented by a circuit or switching table, also called a table of values, one-to-one in the sense that the encoding can be reproduced. In safety-critical circuits it is necessary not defined results, which often occur in complex circuits, to avoid or to monitor. The transferability of circuits into additional and other display possibilities is therefore a necessary property to ensure the functional safety of safety-critical circuits. In this work, a representative combinational circuit is visualized in various ways. In all these representations, however, it should be noted that the "structure-preserving modeling and transfer" is maintained. This means that the formally derived function must consistently match the function derived from the respective representation type. Both functions must in no case have inconsistencies, since only the fault-free function is included in the circuit. Functional safety can be guaranteed by the condition of the *structure-based modeling and transfer.*
To present the application we use an electrical circuit as a use case. And visualize it in various ways like Gate Level (GL), Signal Flow Graph (SFG), Signal Flow Plan (SFP), Module View (MV), Resolution Method (RM) and KV diagram. During creation of different display possibilities, we explain the rules of the structure-based modeling and transfer. In addition, the mathematical axioms, which are based on Propositional Logic (AA), are declared. The advantage of the method is that each type of representation (respectively presentation) has a depth of accuracy, clarity and compactness. The transferability of circuits into other possibilities of directed representation is a necessary property to ensure the functional safety of safety critical circuits.

Organization of the paper: First, the theoretical foundations are briefly explained in Chapter 2. They are regarded as basic knowledge in order to understand this work. Subsequently, the implementation is described in detail by an example in Chapter 3 and visualized by sketches and mod-

*Corresponding Author: Feim Ridvan Rasim, feim.rasim@fau.de

https://dx.doi.org/10.25046/aj030158

els. In the end, the results and the core outline of the work are summarized again and an outlook is given.

2 Theoretical Foundations

2.1 Category level (CL)

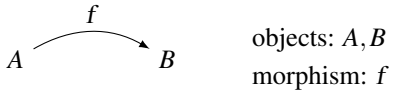


Figure 1: Morphism f from A to B

Morphisms are concretized on category level. A morphism is a directed association between two objects being associative and identitive. In Fig. 1, the morphism f acts on object A and is substituted (=) by object B.

2.2 Value table (VT)

The table of values, also known as truth table or switching sequence table, is a tabular list of the truth value course of a logical statement. In this table, the assignments of the inputs are linked together and the states at the outputs are shown binary. Value Tables (VT) are used for logic operations such as AND, OR, NAND and NOR gates, but also for flip-flops or complex circuits [3], [4]. A value table thus serves to represent the value of a composite statement as a function of the truth values of its partial statements. Tab. 1 shows a truth table for a NOT (left), AND2 (middle) and OR2 (right) function. The input for a NOT function is A and for AND2 and NOR2 the inputs are A and B.

A	NOT A	A	B	A AND B	A	B	A OR B
0	1	0	0	0	0	0	0
1	0	0	1	0	0	1	1
		1	0	0	1	0	1
		1	1	1	1	1	1

Table 1: Value table for NOT (left), AND2 (middle) and OR2 (right) function in (0,1)

2.3 Combinational circuits (CC)

A	B	A NOR B
0	0	1
0	1	0
1	0	0
1	1	0

Table 2: Value table for NOR2 circuit

Under a combinational circuit is a circuit that is realized with simple basic gates such as AND, OR and inverters to understand. It realizes a one-to-one mapping, a function. The outputs of this logic are dependent on the inputs, which means there is no feedback from the outputs to the inputs. The output variable is thus only a function of the input variables [4]. Tab. 2 shows the value table for NOR2 circuit. NOR2 represents a combinational circuit, because the circuit does not include feedback.

2.3.1 Abstraction of a circuit

Under an abstraction of a circuit is a kind of "simplified" representation to understand. A complex, combinational circuit is shown in a different way, in order to make it easier to understand. However, in the various representations, the core message of the output circuit must not be changed. So the function of the circuit must not be changed when transferring it to another representation. The circuit must deliver the same function, no matter in which visual presentation it is shown. For example, an abstraction of a circuit is the transfer of a circuit from the transistor level to the gate level. In addition, a circuit can be presented in its module view (MV) (see section 2.10), which is an isomorphism to the signal flow plan of the circuit. The aim of an abstraction of a circuit is the simplified or clearer visual design or presentation.

2.3.2 Structural changeover and modeling

In order to be able to carry out a structure-faithful modeling, it is to be known that two parallel connected transistors are combined in propositional logic as follows. Fig. 2 shows two transistors connected in parallel. V_{DD} is the operating voltage.
It is a complex gate with two inputs $(\bar{A}, \bar{B})$ and one output (C) between which the logical link "OR" exists. This OR2 outputs "1" at the output when one of the inputs are assigned a "0". This means that if one of the two inputs is assigned a "0", the output creates a "1". The two transistors are connected in parallel, they must be concatenated (addition in field algebra). Furthermore the operating voltage V_{DD} has to be considered. It runs in series with each of the two transistors, it muss be catenated (order, multiplication in field algebra) with two transistors.

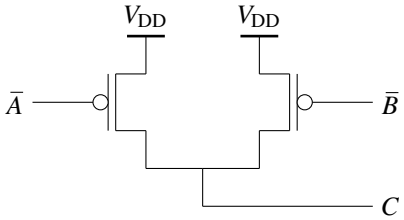


Figure 2: Example of parallel connected transistors

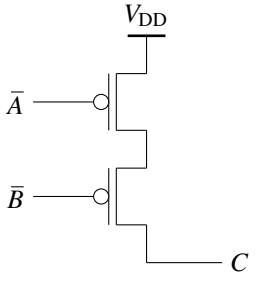


Figure 3: Example of serial connected transistors

If C in Fig. 2 is to be expressed as a function (contrary to logic, only true values can be processed in propositional logic (see section 2.9)), then this is:

$$C = V_{\mathrm{DD}} \cdot (\bar{A} + \bar{B}) \tag{1}$$

In Fig. 3 two transistors are connected in series. It is a complex gate with two inputs $(\bar{A}, \bar{B})$ and one output (C) between which the logical link "AND" exists. This AND2 outputs "1" at the output when both inputs are assigned a "0". This means that if both of the two inputs is assigned a "0", the output creates a "1". Furthermore the operating voltage V_{DD} has to be considered. It runs in series with each of the two transistors, it muss be catenated with two transistors. C can be expressed in propositional logic as follows:

$$C = V_{\mathrm{DD}} \cdot (\bar{A} \cdot \bar{B}) \tag{2}$$

Structure-faithful modeling unites function and structure one-by-one in the sense of a monomorphism injective - that is, the structure has at most one solution (this is the function) - and of a epimorphism surjective - that is, the function has at least one solution (that is the structure). Such a mapping enables a one-to-one (local-bijective) and understandable description of a generating system. During transferring into various presentation possibilities the structure-faithful modeling has a significant role. It is extraordinary important that the formally derived (modeled) function coincide with the function generated by the real structure. Consequently the function has to correspond to reality and shall not exhibit any inconsistencies. Only in this way the functionality of a circuit can be ensured. Structure-faithful therefore means that the relation to reality must never be lost during modeling. Shortly spoken, each pin of the model at gate level (GL) must show up in the real world at transistor level (TL). But, pins have to be correctly labelled in reality at transistor level (TL). This is mandatory.

During the transfer, it is also important that the function generated from the real structure consistently matches the function derived from the signal flow graph or any other type of presentation. Only in this way the functionality of the generating circuit can be ensured. In addition, there is a structure-based transfer only in the absence of inconsistencies. A transfer of the signal flow graph or of the circuit into a value table must also be structure-faithful. Thus the function derived from the evaluation table must correspond to the same function derived from the signal flow graph or the generated circuit. Shortly spoken, each undefined pin of a given reality or not proper assigned signal of a given model has to be shifted to undefined. We consistently use the symbol "$*$".

2.4 RS Buffer

In complex circuits, many structures exist that can create undefined results. These undefined results must not occur in safety-critical circuits, since otherwise the desired function of the circuit cannot be guaranteed. For this reason, the RS buffer structure is established [2]. It can intercept undefined cases in combination with a dual-rail approach. Thus, it is possible to stabilize a complex circuit in its function without glitch. These stabilized states do not produce unpredictable events and can therefore be processed by the circuit without causing errors. Fig. 4 shows the circuit symbol and Fig. 5 shows the circuitry of the RS buffer. The value at the node X in the circuit corresponds to the value at the pin $\bar{Y}$, because of the inverter. Thus, the X is neglected for the sake of clarity in the VT of Tab. 3. On closer examination of Tab. 3, it is noticeable that the RS buffer triggers a switching process only during assignments $(S, \bar{R}) = (1,0)$ and $(S, \bar{R}) = (0,1)$. The old state is retained for assignments $(S, \bar{R}) = (1,1)$ and $(S, \bar{R}) = (0,0)$. The function for the output Y is therefore $Y = R(S \vee Y) \vee S(R \vee Y) = YR \vee YS \vee SR$.

S	$\bar{R}$	Y
0	0	Y
0	1	0
1	0	1
1	1	Y

Table 3: Value table of the RS buffer [2]

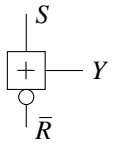


Figure 4: Circuit symbol of the RS buffer

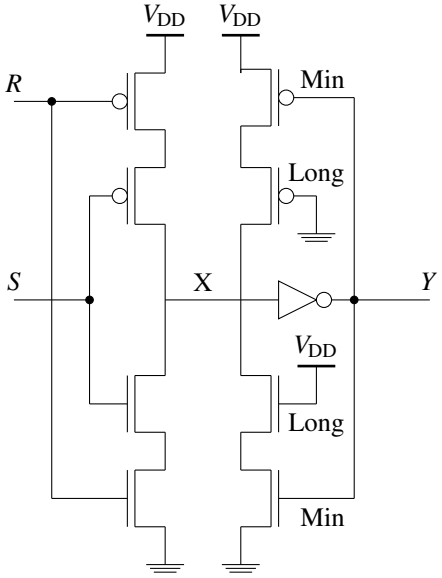


Figure 5: Circuit of the RS buffer [2]

2.5 Signal flow graph (SFG)

The signal flow graph (SFG) is a vividly method to present the internal structure of a system or the interaction of several systems. This presentation allows a better understanding of the function as well as the interrelations of one or more systems. In addition, the signal flow graph is the appropriate tool for abstracting functions or connections to one level above category level (associativity and not identity). The signal flow graph is a directed and weighted graph whose nodes represent objects (sets) and its edges morphisms (functions). The edges of this graph can be understood in a dual view (SFP) as small processing units which process incoming signals (edges) in a particular form and then send the result to all outgoing edges (signals). Signal

flow graphs are formally defined graphs [5]-[9].

2.6 Signal flow plan (SFP)

The signal flow plan (SFP) has a special significance in control engineering and, with the representation method, is based on a block diagram and adds to a further loan of the relationships within a system. A signal flow plan is used to identify the complexity of a system. This contains unidirectional blocks, also called nodes, which transmit incoming signals as small processing units into outgoing signals. In the signal flow graph these are represented by edges. From an SFP, the transfer function of a system can be derived. The signals are redirected to edges, unlike the signal flow graph. In the signal flow graph, they are represented by nodes. Edges in a signal flow plan are also directed connections between two nodes. They illustrate the effect of a signal by its weighting. Furthermore, it is possible to transfer a signal flow graph into a signal flow plan by transferring nodes into edges and edges into nodes [5], [7], [9]. Fig. 6 shows a simple example of a signal flow plan derived from a signal flow graph. The input signals x_0 and x_1 adds up to an output signal y.

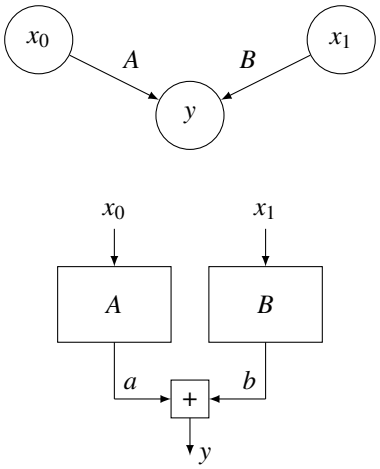


Figure 6: From a signal flow graph (above) to a signal flow plan (below)

2.7 Boolean algebra (BA)

The switching algebra (boolean algebra) is based on decisions and comparisons, so it can explain and visualize logical links very well. Successful results are represented by a "1", unsuccessful results by a "0". These two symbols are complementary to one another. At any time, each pin must be occupied, because only then the system can be a total system and can be calculated by the switching algebra. The switching algebra is not sufficient for a detailed representation of a circuit. However, it is suitable for the functional description without restrictions to the general [10].

2.8 Positive logic (PL)

In the positive logic the symbol "1" stands for a successful event. Unsuccessful events are called undefined. Positive logic is an event that occurs just as it is expected. This means that if a negative event is expected and it occurs, this event is considered successful. This also applies analogously to a positively expected event. If a positive event is expected and it occurs, then this event is also successful. In the positive logic, therefore, only the "1" exists as a value [11]. Here, we need a "0", too. Therefore, the "0" is also a "1" but only a part (child) of the "1".

2.9 Propositional logic (AA)

The propositional logic comes from the formal logic and can be continued into the switching algebra without restrictions to the general. This "Aussagenlogischer Ausdruck (AA)" describes the relationship between statements. Statements can be seen partially in the propositional logic. This means that there should be only one unary statement. The symbols are true (w) and not false ($\bar{\text{f}}$). The following example is intended to illustrate the propositional logic: The statement $Y = (A \wedge \bar{B}) \vee C$ contains two statements and is nevertheless unary, regardless whether is equal to a positive literal or a negative literal. The following statement contains only one statement and is also unary $\bar{Y} = (A \wedge B)$. Thus, in AA a statement is always true, only its complemented content can be interpreted as "false". The logical sign for a true statement is "w " or "$\bar{f}$" [11], [12].

2.10 Module view (MV)

Through a model representation, a real system can be displayed simpler and clearer. The natural section of reality can be simplified by model visualization, and the model representation can also serve as a reference model for further development of a system. In order to represent a real circuit with real electrical components as a model, the structure-faithful transfer is of great importance. In addition, it is necessary that in safety-critical circuits, the modeling is performed structurally, because only then the structure-related functional safety of the circuit is guaranteed. This means that the formally derived and modeled function with which the function created by the real structure must be absolutely identical. Fig. 7 shows a simple example of the asymmetric and irreflexive module view. Due to the directed representation as a module view, a clear visualization is achieved. You can see the input (input vector) X, the output F and the programming vectors Z and Y. By presenting it as a module view, the complexity of the representation can be completely directed.

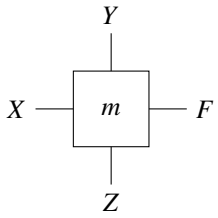


Figure 7: Example of a module view

Only the inputs, outputs and programming vectors are seen. The function, how exactly the interaction of these variables ever is, is not described and explained. The function (transitive closure) is represented by m (model) [5].

2.11 Resolution method (RM)

The resolution method (RM) is a method to check the truth value. It is primarily about being able to prove the satisfiability or unfulfillability. In order to provide this proof, a clause set is extended by new clauses, called resolvents, until an empty clause is generated. If this succeeds, the initial quantity is unsatisfiable. Here, the clauses are conjunctions. A set of clauses, therefore, is a disjunction. It follows that if an empty clause can generated, the initial set is tautological. Resolvent of clauses C_1 and C_2 (according to literal l) is given in [13]. An example of the resolution method is shown in Fig. 8. Here you can see that the clause set can be non-tautologically fulfilled, since in the end no empty clause can be generated. Will be a clause set with the resolvent resulting from their clauses, this results in a logically equivalent set of clauses. This can be reunited with its resolvents without affecting satisfiability, and so on. If finally the empty clause can be formed, a tautological set is proven. In the other case, for example if the last clause is not empty, this is the proof of the satisfiability of the disjunvtive clauses.

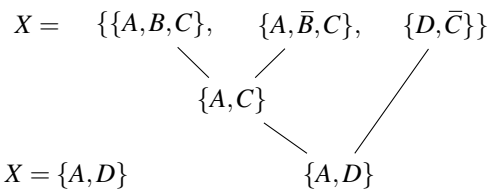


Figure 8: Example of a resolution

The resolution method obeys the following neighborhood relations:

- The resolver $\{y,z\}$ (largest common cover) can be generated from the clauses $\{x,y\}$ and $\{\bar{x},y,z\}$.
- The resolver $\{\}$ can be generated from the clauses $\{x\}$ and $\{\bar{x}\}$.

Definition: Let C_1 be a clause containing the literal l and let C_2 be a clause containing the literal $\bar{l}$. Then the clause is called:

$$C = (C_1 \backslash \{l\}) \cup (C_2 \backslash \{\bar{l}\}) \tag{3}$$

2.12 KV diagram

The Karnaugh-Veitch diagram (KV diagram) is used for the clear presentation and simplification of Boolean functions into a minimal expression. A KV diagram can be used to transform any disjunctive normal form (DNF) into a minimal disjunctive logical expression. The diagram is labeled with the variables at the edges. Each variable occurs in negated form (negative literal) and not negated form (positive literal). The assignment is arbitrary. However, it should be noted that horizontally and vertically adjacent fields may only differ in exactly one variable [4].

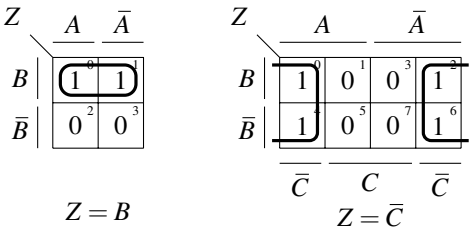


Figure 9: KV diagram examples

Minterm Method:

- Try to group as many horizontally and vertically adjacent fields as possible that contain a "1" (1, 2, 4, 8, 16, 32, ...).
- A block may continue over the right / left or bottom of the chart.
- From these identified blocks so many are to be selected that all "1" fields are covered (Fig. 9).
- Now the conjunct terms are formed.
- These conjunctive terms are linked with an OR and a DNF results.

Maxterm method: This method differs from the Minterm method in the following points:

- Instead of ones, zeros are combined into blocks.
- Subsequently, disjunction terms are formed.
- These disjunction terms are AND, resulting in a conjunctive normal form (CNF).

3 Implementation

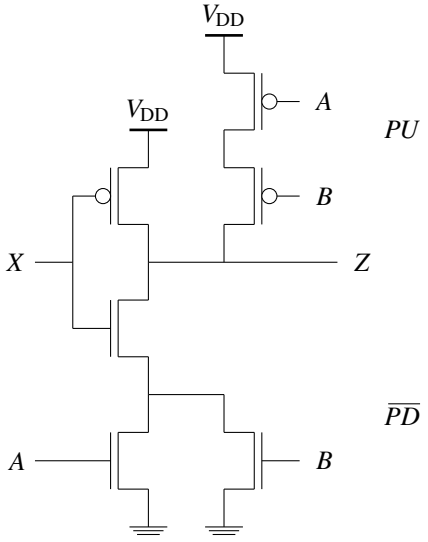


Figure 10: Simple electrical circuit at transistor level (TL)

In this Chapter, an electrical circuit is considered at the transistor level in Fig. 10. Afterwards, the circuit is transfered from the transistor level into the gate level in a structured manner. It should be noted that the analog circuit, that

is, the circuit at transistor level, is described at the gate level in propositional logic. Subsequently, the circuit is converted into a signal flow graph and signal flow plan at the gate level. Various possibilities for the representation of the output circuit, then such as the evaluation table, the module view or the resolves are presented. With all these possibilities of representation, it should be noted that the respective derived function must not have any inconsistencies that means, the formally derived function must agree with a function generated from a real structure (TL). Fig. 10 shows the circuit at the transistor level, this circuitry is a combinational circuit because there are no feedback lines, which is to be transferred to the gate level. It is a complex gate with three inputs and an output.

3.1 Concretization to gate level

The analysis of a circuit at the transistor level is more detailed and more complex than viewing at the gate level, since the representation in gates is only a "model" which allows a simplified and clear view of the circuitry. The transfer of a structure at the transistor level into a structure at the gate level is therefore a concretization (often called abstraction) and serves to increase the clarity and simplify the understanding of the structure. From propositional logic and category point of view, however, transistor level (parent) is the abstraction of the gate level (child). This is important to keep in mind.

In the first step is now the electrical circuit transferred to the gate level. First, the pull-up (PU) and the pull-down $(\overline{PD})$ are viewed separately from each other. When the transistors of the pull-up or pull-down are connected in parallel, they must be concatenated. On the other hand, when the transistors are connected in series, they must be catenated. Furthermore the operating voltage V_{DD} has to be considered in the pull-up. This runs in series with each of the two lines of the pull-up. And the mass potential runs serially to both line of the pull-down.
The transistors in the *PU* are switched with a "0" low-active input and the transistors in the $\overline{PD}$ are switched with a "1" high-active input. *PU* and $\overline{PD}$ are connected by a composition (concatenation). With the composition it is meant that different things are substituted but do preserve their directed manner (the morphisms are preserved). This composition ensures the RS buffer in Fig. 5. In Fig. 5 after the pull-down, a switch "¬" is installed in order to meet the propositional composition (we use switch to replace the two levels "0" and "1", we can switch from one level to another). The function of the RS buffer has already been explained in the basic chapter. The operating voltage is usually designated with V_{DD}, the reference point is the mass with the low-active input $\overline{\mathrm{GND}}$ (since the gate level has been in AA, the ground is written as $\overline{\mathrm{GND}}$). Please keep in mind that pins have to be coded correctly. You have to choose the correct literals, either positive literals or negative literals.

In the second step subcircuits are described as concrete mathematical functions in the propositional domain (in the AA domain). It should also be noted that each partial circuit is basically at least disjunct, in this case even complementary, this mean a pull-up (PU) and a pull-down $(\overline{PD})$. Pull-up means that the output is pulled up to the operating voltage V_{DD}. This part of the circuit is low-active since a logical "0" must be present at the primary input in order to trigger the switching process. The $\overline{PD}$ draws the output to the mass potential. It is referred to as high-active, since a logical "1" must be present at the primary input, so that a switching process is triggered. For example, Tab. 3 shows the functionality of the RS buffer. The operation is assumed to be already known. Thus, the circuit consisting of a *PU* and a $\overline{PD}$ has the following equation:

$$Y = PU + \neg\overline{PD} = S + \neg\bar{R} \tag{4}$$

Eq. 4 will play an important role in the later course of the work. After all steps, as explained above, have been carried out, a structure-preserving model at the gate level results. Important is, that during all steps inconsistencies must not occur. In Fig. 11, the electrical circuit is now displayed in propositional logic at gate level. As described above, the *PU* and $\overline{PD}$ are summarized using the composition. In addition, a switch is built between $\overline{PD}$ and the composition. Its task is also to highlight the low-active input of the RS buffer. It is important that the stars here in the continuation only serve as a "monitor" for checking. If the circuit is designed correctly, each star $(*)$ can only supply a "0".

In the third step, the node *Z* (pin *Z*) is expressed in a function: The *PU* and $\overline{PD}$ is connected to the composition "+". The operating voltage V_{DD} flows in series with the inputs in the *PU*. The mass runs serially to all inputs in the $\overline{PD}$. The switch between composition and pull-down switches the logic value of the last part to a negative literal - the substitute of its complement in Eq. 4. In summary, it should be emphasized that viewing at the gate level (in propositional logic) allows a directed and more simplified view, which makes the derivation of the function at node *Z* easier.
The node *X* is treated as another input. Thus, it must be noted that node *X* affects the *PU* and the $\overline{PD}$. Furthermore, the *PU* and the $\overline{PD}$ are linked by a composition. This composition works like an RS buffer. The inputs are then linked to the logical operators. Not to forget is the switch, which is between $\overline{PD}$ and the composition. This "switches" the part of the function belonging to the $\overline{PD}$. The operating voltage V_{DD} and the mass run serial to the inputs.
The stars $(*)$ also serve (Fig. 11) for to check the correctness of the circuit. For example, each individual star can only assume the value "0" in order to ensure that the circuit is correctness. Fig. 11 shows the structure-faithful modeling of the circuit at gate level (in AA). After the circuit is transferred to gate level, the node *Z* can be expressed as a mathematical function in Eq. 5.

$$Z = V_{\mathrm{DD}} \cdot (\overline{X} + \bar{A} \cdot \bar{B}) + \neg(\overline{\mathrm{GND}} \cdot (A + B) \cdot X) \tag{5}$$

The circuit at gate level in Fig. 11 shows the concreteness of the output circuit (Fig. 10). As described above, it was transferred from the transistor level to the gate level in a structure-preserving manner. The functions derived at the transistor level and the functions derived at the gate level must be identical with respect to their partial order (directed manner), this means the function of the circuit must

not be changed by the transfer. Only then is the transfer a structure-faithful one.

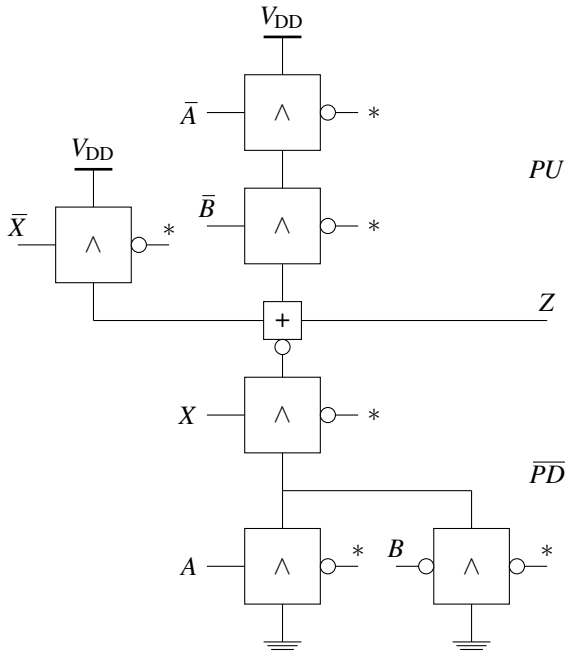


Figure 11: Simple electrical circuit at gate level (in AA)

The gate level can be viewed as a model view. It serves to increase clarity as well as contribute to an understanding of the circuit.

3.2 Transfer to a signal flow graph (SFG)

With the basic knowledge from Chapter 2, the function Z is now transferred step by step into an SFG at Fig. 12.

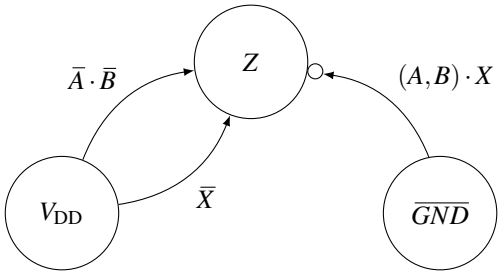


Figure 12: SFG of electrical circuit

For the creation of SFG, it is determined which signals are visualized in a node and which are represented by an edge. The operating voltage V_{DD} and the ground $\overline{GND}$ are visualized by a node, while the inputs, also the X, are used as edge weights for the processing of the signals. The operating voltage is catenated by the inputs $\overline{A}$, $\overline{B}$ and $\overline{X}$. Please remember that we are in propositional logic. The negative literal $\overline{A}$ is not the complement of A but can be substituted by the complement of A, $\overline{A} = \bar{A}$. BA is constituted by $\overline{A} = \neg A$. Thus, V_{DD} represents a node while the inputs with the switch reside on the edges. The third edge has as an edge weight the catenation of the X with the two inputs A and B. The mass is represented as a node. The switch (bubble at node Z) must be considered as shown in Fig. 12. These three edges end at to the node Z. The switch, which must be installed here, is not to be forgotten. These two edges concatenate to the node X. When looking at the Z function it is noticeable that the node X has a major influence on this function:

It is important that the function Z generated from the real structure is consistent (in the direction) with the functions derived at the gate level and the signal flow graph. The SFG allows a further comprehensible and simple visual consideration of the problem. The system is represented simply and visually by weighted, directed graphs. In dual sense (SFP), blocks are small processing units that process incoming signals (that are edges) in a certain form and send the result to all outgoing edges (that are signals). In Fig. 12, the comma "," means in parallel. We use the comma "," to write different things next to each other but preserving the directed manner. Since the directions have to be preserved in the structure faithful modeling and transfer, we announce two edges A and B next to each other. From Fig. 12 follows equation 6:

$$Z = V_{DD} \cdot (\overline{X} + \bar{A} \cdot \bar{B}) + \neg(\overline{GND} \cdot (A + B) \cdot X) \qquad (6)$$

3.3 Transfer to a signal flow plan (SFP)

As already explained, the node can be interpreted in the dual sense as a partition, a signal, and an edge over its weight as processing (operation) of the signal. Thus it generates (substitutes) a new signal. The states of the output circuit are to be found in the nodes. The edges are supplemented with their weights. If the electrical circuit is now considered more closely at gate level, the background knowledge of this work can be used to derive a signal flow plan. It is important to know that the function which is derived from the signal flow graph has to coincide with the function which is derived at the gate level. For only then the modeling has been done in a structured way and the relation to reality has not been lost. In order to be able to transmit a signal flow graph into a signal flow plan, the following must be carried out. By interchanging nodes and edges, a signal flow graph results in a signal flow plan and vice versa.

In order to set up a signal flow plan one has to determine which signals are visualized in a node and which are represented by an edge. The edge is a directed line, which connects two nodes and effects the processing of a signal via its weight. The mass as well as the operating voltage represent the nodes in the SFG. The primary inputs are shown as edges. The signal flow plan (action plan) is used to determine the complexity of a system. The nodes (blocks) in a signal flow plan are small processing units that transmit incoming signals on edges into outgoing signals on edges. Fig. 13 shows the signal flow plan derived from the signal flow graph of the electrical circuit. The signals, in this case the edges (nets) of the signal flow plan, are found on the nodes of the signal flow graph. The nodes (blocks) of the signal flow plan are found in the edges of the signal flow graph.

After the gate level of node Z has been transferred to a signal flow plan, the signal flow plan is displayed in a module view (Fig. 14).

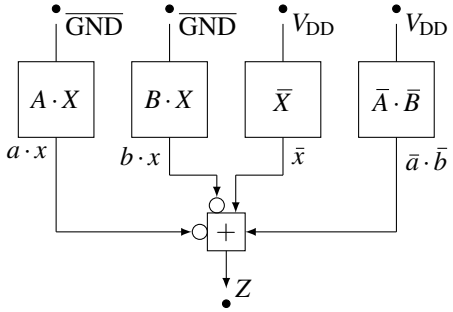


Figure 13: SFP of an electrical circuit

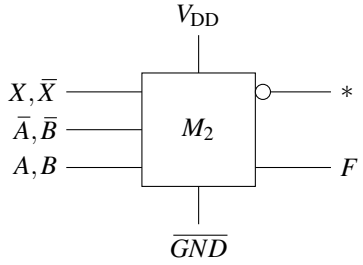


Figure 14: Module-view of an electrical circuit

This digital structural element serves for further simplified viewing of the real system. This further simplifies understanding of the structure. The module-view for node Z has:

- an input vector (inputs): A, B, $\overline{A}$, $\overline{B}$ and $X, \overline{X}$
- a programming vector (states): $\overline{\text{GND}}$ and V_{DD}
- an output vector (output): $Z =: F$

As the MV is using the transitive closure it preserves all pins (that occur in reality). Therefore, it is structure-faithful and a compact presentation of the circuit.
Within M_2, the function is outdated here. We can convert this function from MV to other display posibilities structure-faithful.
At the star $(*)$ we should only observe the "0".

3.4 Evaluation table of the circuit

In the next step the node Z (Eq. 7) of the output circuit (Fig. 11) is shown in the form of a total switching table (in positive logic (Tab. 4), followed by the partial switching table in positive logic (Tab. 5).

$$Z = V_{\text{DD}} \cdot (\overline{X} + \overline{A} \cdot \overline{B}) + \neg(\overline{\text{GND}} \cdot (A + B) \cdot X) \qquad (7)$$

The function in Eq. 7 can be expressed as follows:

$$\begin{aligned} Z &= (Z, \overline{Z}) \\ &= V_{\text{DD}} \cdot (\overline{X} + \overline{A} \cdot \overline{B}), \neg(\overline{\text{GND}} \cdot (A + B) \cdot X) \\ &= V_{\text{DD}} \cdot (\overline{X} + \overline{A} \cdot \overline{B}) + \neg(\overline{\text{GND}} \cdot (A + B) \cdot X) \\ &= PU + \neg\overline{PD} \end{aligned} \qquad (8)$$

In Eq. 7, the node Z is represented in different versions. $Z = (Z, \overline{Z})$ means that Z is a partition of two blocks provided by a positive and a negative literal. The last part of Eq. 7 means that Z consists of two parts, the pull up and the pull down. The relationship between PU and $\overline{PD}$ is the same as already shown in Eq. 4, not to forget the composition "+" between PU and $\overline{PD}$, the RS buffer.
Tab. 3 is also required in order to be able to set up the switching table. The operating voltage V_{DD} depends on the pull-up. It is important to know that if only the operating voltage supplies a "1", a reliable switching process can be present in the PU. The pull-down depends on the mass, $\overline{\text{GND}}$. A switching operation in the $\overline{PD}$ can only take place when the mass is at "0".
Tab. 4 shows the total switching table of node Z in positive logic (PL). On closer examination of the table it can be seen that this table is composed of three divisional tables. The result of the PU and the $\overline{PD}$ and the node Z represent a further table. The coding universe for the PU consists of the operating voltage V_{DD} and $(\overline{A}, \overline{B}, \overline{X})$. For the $\overline{PD}$, the coding universe consists of (A, B, X) and $\overline{\text{GND}}$. Thus the table for the PU and $\overline{PD}$ has a total of $2^4 = 16$ assignments. The output Z represents the respective state for the one-to-one coding of the table.
The lower part of the table (last eight lines) represents assignments which tend not to trigger any switching operations. The reason for this is that during the pull-up the operating voltage V_{DD} can only assume the value "1". The mass does not matter. For pull-down, $\overline{\text{GND}}$ must only have the value "0", so that a switching operation is triggered. The operating voltage does not matter.
The results of the pull-up and the pull-down are calculated by Eq. 7. In order to determine the resulting node Z, Tab. 3 is to be considered. It represents the relationship between PU and $\overline{PD}$.
In the next step, the total switching table Tab. 4 (in PL) is transferred to a partial switching table Tab. 5 (in PL). The correctness of the partial switching table is only given because a structure correct transformation and modeling has taken place. This ensures that the circuit has been designed without errors and thus a partial switching table can be applied without errors.
Tab. 5 shows the partial switching table of node Z. The operating voltage is unsignificant for the $\overline{PD}$, whereas in the PU it can trigger a switching operation only with a "1". The mass is meaningless for the PU, whereas only a "0" the mass in the $\overline{PD}$, triggers a switching process. Thus the last eight lines of the total switching table is lost. For the safety of a defect-free structure it can be firmly assumed that in these assignments, the output Z assumes the states in Tab. 5.
The symbol $*$ represents undefined.

3.5 Resolution method for the circuit

In the last step, the node Z (Eq. 9) of the output circuit (Fig. 11) is entered into a KV diagram.

$$Z = V_{\text{DD}} \cdot (\overline{X} + \overline{A} \cdot \overline{B}) + \neg(\overline{\text{GND}} \cdot (A + B) \cdot X) \qquad (9)$$

PU					$\overline{PD}$							
$\overline{\text{GND}}$	$\overline{X}$	$\overline{A}$	$\overline{B}$	V_{DD}	V_{DD}	X	A	B	$\overline{\text{GND}}$	PU	$\overline{PD}$	Z
$*$	0	0	0	1	$*$	0	0	0	0	1	$*$	1
$*$	0	0	1	1	$*$	0	0	1	0	1	$*$	1
$*$	0	1	0	1	$*$	0	1	0	0	1	$*$	1
$*$	0	1	1	1	$*$	0	1	1	0	1	$*$	1
$*$	1	0	0	1	$*$	1	0	0	0	1	$*$	1
$*$	1	0	1	1	$*$	1	0	1	0	$*$	1	0
$*$	1	1	0	1	$*$	1	1	0	0	$*$	1	0
$*$	1	1	1	1	$*$	1	1	1	0	$*$	1	0
$*$	0	0	0	0	$*$	0	0	0	1	$*$	$*$	$*$
$*$	0	0	1	0	$*$	0	0	1	1	$*$	$*$	$*$
$*$	0	1	0	0	$*$	0	1	0	1	$*$	$*$	$*$
$*$	0	1	1	0	$*$	0	1	1	1	$*$	$*$	$*$
$*$	1	0	0	0	$*$	1	0	0	1	$*$	$*$	$*$
$*$	1	0	1	0	$*$	1	0	1	1	$*$	$*$	$*$
$*$	1	1	0	0	$*$	1	1	0	1	$*$	$*$	$*$
$*$	1	1	1	0	$*$	1	1	1	1	$*$	$*$	$*$

Table 4: Total switching table (in PL): PU, $\overline{PD}$ and Z

$\overline{X}$ for PU X for $\overline{PD}$	$\overline{A}$ for PU A for $\overline{PD}$	$\overline{B}$ for PU B for $\overline{PD}$	V_{DD}	$\overline{\text{GND}}$	PU	$\overline{PD}$	Z
0	0	0	1	0	1	$*$	1
0	0	1	1	0	1	$*$	1
0	1	0	1	0	1	$*$	1
0	1	1	1	0	1	$*$	1
1	0	0	1	0	1	$*$	1
1	0	1	1	0	$*$	1	0
1	1	0	1	0	$*$	1	0
1	1	1	1	0	$*$	1	0

Table 5: Partial switching table (in PL): PU, $\overline{PD}$ and Z

For this purpose, Eq. 9 must be converted into a DNF system, as can be seen in the following.

$$Z = V_{\text{DD}} \cdot \overline{X} + V_{\text{DD}} \cdot \overline{A} \cdot \overline{B} + \neg(\overline{\text{GND}} \cdot A \cdot X) + \neg(\overline{\text{GND}} \cdot B \cdot X) \quad (10)$$

Due to the different terms of the Eq. 10, a better understanding of the KV diagram is to be developed in Fig. 15. It is noticeable that some fields overlap in the diagram. In particular, this fields are occupied by "1" and "¬1". The question is how this field actually looks in the KV diagram in AA. Now Eq. 10 is a concrete equation (Category Level). The KV diagram for this equation is to be represented in AA. In this case, the switch "¬" is substituted by complement since the KV diagram is in the Aussagenlogik (AA), the propositional logic, the logic of statements. The KV diagram now looks as Fig. 16. The blocks of the KV diagram (Fig. 16) are now determined from Eq. 10 as follows:

$$(\overline{A} \wedge \overline{B} \wedge V_{\text{DD}}) \vee (X \wedge A \wedge \overline{\text{GND}}) \vee (B \wedge X \wedge \overline{\text{GND}}) \vee (\overline{X} \wedge V_{\text{DD}}) \vee (V_{\text{DD}} \wedge \overline{\text{GND}}) \quad (11)$$

Therefore in Eq. 11, $V_{\text{DD}} \wedge \overline{\text{GND}}$ represents a redundant block, a redundant prime implicant. This case is visualized by dashed lines. This part of the equation is now used for resolving. Fig. 17 shows the resolution method for node Z. The clauses ($\{\overline{A}, \overline{B}, V_{DD}\}$, $\{X, A, \overline{GND}\}$, $\{X, B, \overline{GND}\}$, $\{\overline{X}, V_{DD}\}$) are derived from Eq. 10. In order to be able to determine the clauses, the equation was first transformed into a DNF system in order to connect the switch neglected. Only by neglecting the switch can a resolvent be formed. In order to determine the folowing clauses ($\{\overline{A}, \overline{B}, V_{DD}, \overline{GND}\}$, $\{\overline{A}, \overline{B}, V_{DD}, GND\}$, $\{X, A, GND, \overline{V}_{DD}\}$, $\{X, A, \overline{GND}, V_{DD}\}$, $\{X, B, \overline{GND}, \overline{V}_{DD}\}$, $\{X, B, \overline{GND}, V_{DD}\}$, $\{\overline{X}, V_{DD}, GND\}$, $\{\overline{X}, V_{DD}, \overline{GND}\}$), the ones in the KV diagram (Fig. 16) have to be viewed individually. $\{V_{DD}, \overline{GND}\}$ comes from prime implicant (dashed lines) of Fig. 16.

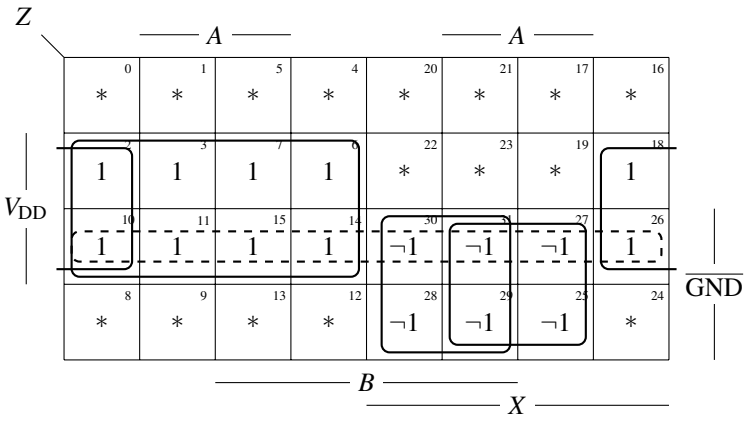


Figure 15: KV diagram for Z (electrical circuit in PL)

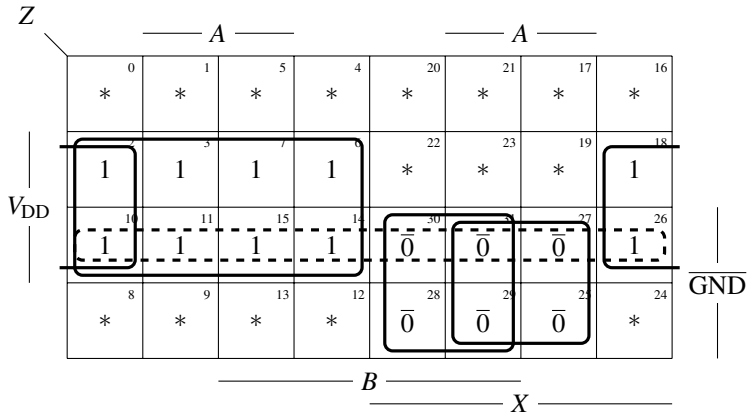


Figure 16: KV diagram for Z (electrical circuit in AA)

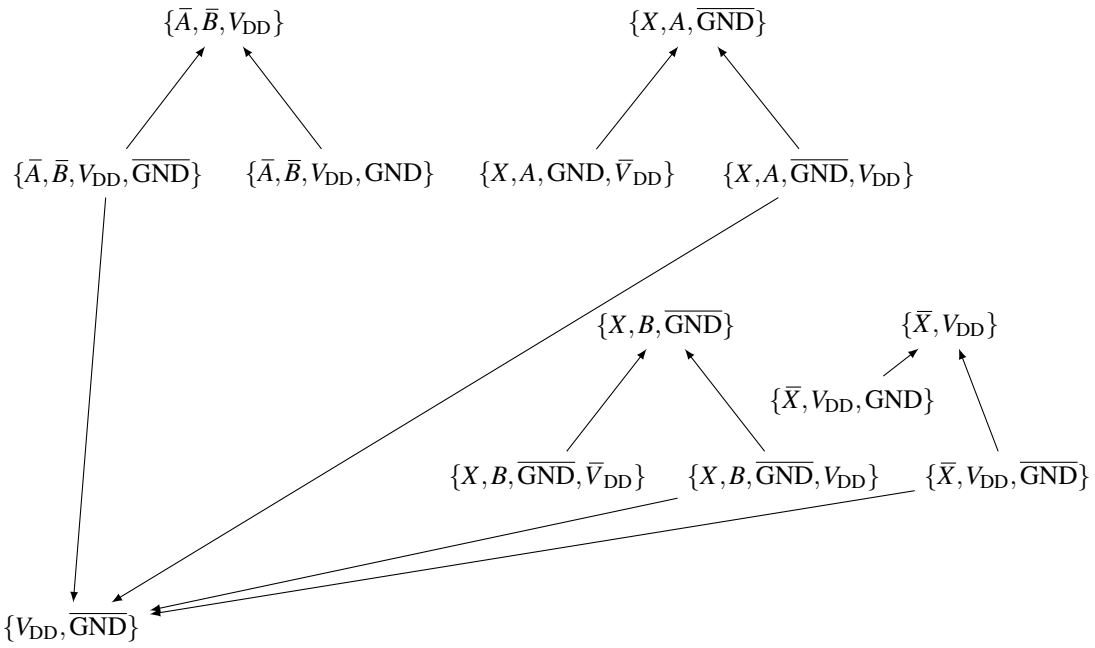


Figure 17: Resolution method for the electrical circuit in PL

So it is noticeable that the KV diagram shows additional information when looking at the ones. The ones that can be useful in resolving are selected. Resolves can be generated only from certain clauses. As already explained in Chapter 2, the following holds: The resolver $\{y,z\}$ can be generated from the clauses $\{x,y\}$ and $\{\bar{x},y,z\}$. The resolver $\{\}$ can be generated from the clauses $\{x\}$ and $\{\bar{x}\}$. The prime implicant (dashed) in the KV diagram is made up of this clause: V_{DD}, $\overline{\mathrm{GND}}$.
This clause resolves from three different clauses, as shown in Fig. 17.

4 Conclusion

The transferability of circuits into other possibilities of representation is a necessary property to ensure the functional safety of safety critical circuits. In this work an output circuit has been visualized in various display possibilities. Each type of presentation has its advantages and disadvantages. Moreover, each type of representation has a depth of accuracy, clarity and compactness. However, all of these representations are common in that their "structure-faithful modeling and transition" must be preserved. This means that a formally derived function has to match consistently with the function derived from the respective representation type. Both functions must in no case have inconsistencies, because only then the fault-free function of the circuit can be maintained. Shortly spoken, each pin of the model at gate level must show up in the real world at transistor level. But, pins have to be correctly labelled in reality. This is mandatory.

References

[1] F. R. Rasim, C. Kocar, S. M. Sattler: Structure-Preserving Modeling of Safety-Critical Combinational Circuits, 20th IEEE International Symposium on DDECS 2017, Dresden

[2] G. Uygur, S. M. Sattler: Structure-Preserving Modeling for Safety Critical Systems. Mixed-Signal Testing Workshop (IMSTW), 2015

[3] M. M. Mano, R. K. Charles: Logic and Computer Design Fundamentals, Third Edition. Prentice Hall, 2004

[4] P. Horowitz, W. Hill: The Art of Electronics, Second Edition. Cambridge Press, 1989

[5] R. C. Dorf, R. H. Bishop, Modern Control Systems Solution Manual, Pearson Studium, 2011

[6] S. J. Mason, Feedback Theory - Further Properties of Signal Flow Graphs, IEEE, vol. 44, pp. 920-926, 1956

[7] W. S. Levine, The Control Handbook, CRC and IEEE Pres, 1996

[8] J. A. Brzozowski, E. J. McCluskey, Signal Flow Graph Techniques for Sequential Circuit State Diagrams. IEEE vol. EC-12, pp. 67-76,1963

[9] L. M. Horowitz, Syntehsis of Feedback Systems, Academic Press INC. London, 2013

[10] R. S. Stankovic, J. Astola: From Boolean Logic to Switching Circuits and Automata, Springer, 2011

[11] L. H. Hackstaff: Systems of Formal Logic, Springer, 2012

[12] Ch. Posthoff, B. Steinbach: Logic Functions and Equations, Springer, 2009

[13] A. Leitsch: The Resolution Calculus, Springer, 1997

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 483-495 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Hardware Acceleration on Cloud Services: The use of Restricted Boltzmann Machines on Handwritten Digits Recognition

Eleni Bougioukou, Nikolaos Toulgaridis, Maria Varsamou, Theodore Antonakopoulos *

Department of Electrical and Computer Engineering, University of Patras, 26504 Rio - Patras, Greece

ARTICLEINFO

Article history:
Received: 30 November, 2017
Accepted: 07 January, 2018
Online: 18 February, 2018

Keywords:
Neural networks
Handwritten Digits Recognition
Cloud servers
Hardware accelerators

ABSTRACT

Cloud computing allows users and enterprises to process their data in high performance servers, thus reducing the need for advanced hardware at the client side. Although local processing is viable in many cases, collecting data from multiple clients and processing them in a server gives the best possible performance in terms of processing rate. In this work, the implementation of a high performance cloud computing engine for recognizing handwritten digits is presented. The engine exploits the benefits of cloud and uses a powerful hardware accelerator in order to classify the images received concurrently from multiple clients. The accelerator implements a number of neural networks, operating in parallel, resulting to a processing rate of more than 10 MImages/sec.

1 Introduction

We live in the era of 'Big Data', where a vast amount of structured, semi-structured and unstructured data are being generated at an ever-accelerating pace and can be mined to obtain valuable information. The most commonly used approach to process this kind of data is to aggregate raw data into large datasets, probably extend them with metadata, and then apply machine learning and/or artificial intelligence algorithms in order to identify repeatable patterns [1]. Artificial neural networks are a rapidly developing category of machine learning structures that give computers the capability to learn without being explicitly programmed to perform specific tasks. They consist of different layers for analyzing and learning data. Each layer consists of a large number of highly interconnected processing elements (neurons), working together to learn from previous data in order to solve specific problems by making proper decisions.

Boltzmann Machine (BM) is a typical example of a neural network structure. BMs are probabilistic Markov Random Field models that use a layer of hidden variables to model a distribution over input variables, called visible variables. In general, learning a Boltzmann machine is a computationally demanding process. However, the learning problem can be simplified by imposing restrictions on the network topology, which leads to Restricted Boltzmann Machines (RBMs)[2]. RBMs are structured as bipartite undirected graphs, which results to efficient inference implementation, and are particularly capable of learning complex features. The last few years, many applications based on RBMs have been developed to cover a large variety of learning problems, such as image classification, speech recognition, collaborative filtering and so on. One such example application is the recognition of handwritten digits. Learning an RBM corresponds to fitting its parameters, so that the distribution represented by the RBM models the distribution underlying the training data, handwritten digits in this case [3]. The storage resources and the time required not only to train an RBM but also to make real-time predictions on new coming data increases exponentially with the number of parameters. Thus the development of a handwritten digit recognition application on a single user machine is a non-trivial task.

Nowadays, cloud computing solutions provide new capabilities to users and enterprises for processing their data remotely in high performance servers. Users do not have to invest in information technology infrastructure, reducing the need for

*Theodore Antonakopoulos, University of Patras, Patras 26504, Greece, Email: antonako@upatras.gr, Tel.: +30(2610)996487

https://dx.doi.org/10.25046/aj030159

advanced hardware at the user side. In addition, cloud providers specialized in a particular area, such as image processing, can bring advanced services to a single user, complex services that are not easily afforded by individual users. Another important benefit is the high processing rate that can be achieved when a high performance server is used for processing requests from multiple independent users.

In this work, which is an extension of the work originally presented in [4] and [5], we exploit the vast amount of resources provided by cloud computing along with the high computation capabilities offered by hardware accelerators in order to build a complete cloud-based engine that can be used in real-time handwritten digits recognition (HDR). The engine collects images from multiple sources over the cloud and processes them as fast as possible resulting in high processing rate. The high processing rate is achieved by using a powerful hardware accelerator that implements a number of neural networks operating in parallel. A major advantage of the proposed engine is that the training of the neural networks is not only performed once during initialization of the system, but it is also fed with new images periodically, thus improving the accuracy of the prediction results. The engine is presented in detail in the sections that follow.

Section 2 gives an overview of existing implementations, especially on handwritten digits recognition. Section 3 analyzes Restricted Boltzmann Machines and how they can be modified in order to be used to solve classification problems, such as image recognition. Section 4 presents the architecture and the functionality of the proposed cloud-based computing server. Section 5 highlights the communication interface between the server and the hardware accelerator. Section 6 describes in detail the implementation of the neural networks, with emphasis on the architecture of the dedicated hardware accelerator. A complete system prototype was developed, which is presented in Section 7. Experimental results that demonstrate the system performance in terms of processing rate for both implementations are also presented.

2 State of The Art

The scientific area of automatic handwriting recognition is of great interest for both academia and industry. Existing algorithms are so efficient in learning to recognize handwritten digits that they are used, for example, by post offices to sort letters and banks to read personal checks. MNIST is the most widely used dataset for studying handwritten digit recognition [6]. State-of-the-art models that present accuracy results in the range of 0.35% down to 0.23% error rates are based on large Convolutional Neural Networks (CNNs), either in the form of a deep single network optimized with various training techniques or as a committee of many smaller networks [7], [8], [9]. The best results so far, 0.21% error rate, have been claimed by an approach based on regularization of neural networks using DropConnect [10].There is a listing of the state-of-the-art results and links to the relevant papers on the MNIST and other datasets collected by Rodrigo Benenson [11].

The disadvantage of CNNs is that they are very resource-demanding and their training and inference procedures are extremely time-consuming. As the amount of data increases, machine learning moves to the cloud and big clusters of high-performance servers are used for providing real-time results. Most works mainly deal with implementing the training procedure by using distributed servers over the cloud ([12], [13]) or by using higly scalable FPGA implementations ([14], [15]). In this work, a high performance computing engine that accepts images from multiple clients over the cloud and classifies them with high accuracy is presented.

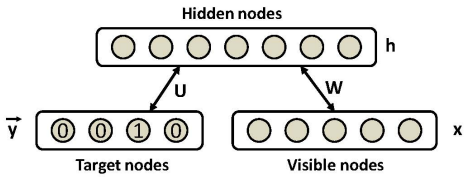


Figure 1: Discriminative RBM modeling the joint distribution of inputs and target classes.

3 Classification Restricted Boltzmann Machines

Restricted Boltzmann Machines are usually used as feature extractors for other learning algorithms or as initializers for deep feedforward neural network classifiers, not as standalone classifiers. However, authors in [16] have proposed a discriminative variant of RBMs that can be used autonomously in classification tasks offering good performance results. The bipartite undirected graph of such an RBM is illustrated in Figure 1. Given a training set $D_{train} = \{(x_1, y_1), ..., (x_i, y_i), ..., (x_D, y_D)\}$, where i denotes the i-th example of the set consisting of an input vector x_i and the corresponding target class $y_i \in \{1, ..., C\}$, we use the specific RBM to model the joint distribution between a layer of N hidden variables $\mathbf{h} = (h_1, ..., h_N)$, usually referred as features, and the observed variables $(\mathbf{x}, \mathbf{y})$. It is a parametric model where the parameters Θ = **(W,b,c,d,U)** represent the following:

- **W:** Weights matrix between **x** and **h**
- **U:** Weights matrix between $\mathbf{e}_y$ and **h**
- **b,c,d:** Respective biases of **x**, **h** and $\mathbf{e}_y$

and $\mathbf{e}_y = (1_{i=y})$ for $i \in \{1, ..., C\}$ the 'one out of C' vector representation of y.

We consider the binary version of the model where each node, hidden or visible, may be in one state, **ON** or **OFF**. A node adopts a new state as a probabilistic

function of the states of its neighboring nodes and the weights on its links to them. It stands:

$$p(h_j = 1|y,x) = sigm(c_j + u_{jy} + \sum_i w_{ji}x_i) \quad (1)$$

$$p(x_i = 1|h) = sigm(b_i + \sum_j w_{ji}h_j) \quad (2)$$

$$p(y|h) = \frac{\exp(d_y + \sum_j u_{jy}h_j)}{\sum_{\text{for all y}} \exp(d_y + \sum_j u_{jy}h_j)} \quad (3)$$

where $sigm(x) = 1/(1 + e^{-x})$ is the logistic sigmoid function.

After the model has been trained, the conditional probability $p(y|x)$ is used for classification. The conditional probability $p(y|x)$, can be computed using $p(y|x) = argmin(F(y,x))$:

$$F(y,h) = -d_y - \sum_j \log(1 + e^{(c_j + u_{jy} + \sum_i w_{ji}x_i)}) \quad (4)$$

where $F(y,h)$ is called *free energy*.

In order to train an RBM to solve a particular classification problem, an objective has to be defined that the learning procedure will try to minimize for all examples in the dataset D_{train}. It is possible to choose among various different objective functions, but generally the following three are used [17]:

- Generative Training Objective
- Discriminative Training Objective
- Hybrid Training Objective

3.1 Generative Training Objective

Given that the model defines a value for the joint probability $p(x,y)$, a natural choice for a training objective is the generative objective:

$$L_{gen}(D_{train}) = -\sum_{i=1}^{|D_{train}|} \log p(y_i, x_i) \quad (5)$$

Computing $\log p(y_i, x_i)$ and its gradient with respect to any RBM parameter Θ is intractable. Fortunately it has been shown that the gradient can be well approximated using the Contrastive Divergence estimator. The analytical computation is replaced by an estimate at a sample generated after a limited number of Gibbs sampling steps, with the sampler's initial state for the visible variables set at the training example (x_i, y_i). A single Gibbs sampling iteration is usually sufficient to learn a meaningful representation of the data [18]. Then, this gradient estimate can be used in a stochastic gradient descent procedure for training. A pseudocode of the procedure is given in Algorithm 1, where γ is the learning rate. Usually, the weights **(W,U)** are initialized using small random values, while the biases **(b,c,d)** are initially zero. Ideally, RBMs require parameter updating after each single example, but mini-batch updating can be also used. However, in order to ensure fast model convergence, the batch size should remain relatively small.

Algorithm 1 RBM Training over (x,y) using 1-step Contrastive Divergence

Input: w_{ij}

for all training samples (x,y) **do**

1. Calculate $p(h = 1|x,y)$ for all hidden nodes
   Sample the hidden distribution $< h_0 >$
2. Perform Gibbs sampling for k steps (k=1):
   Calculate $p(x|h_0)$ for all visible nodes and $p(y|h_0)$ for all target nodes
   Sample the visible distribution $< x_k >$ and the target distribution $< y_k >$
   Calculate $p(h = 1|x_k, y_k)$ for all hidden nodes
   Sample the hidden distribution $< h_k >$
3. Calculate gradients:
   $gW =< h_k > * < x_k > - < h_0 > *x$
   $gU =< h_k > * < y_k > - < h_0 > *y$
   $gb =< x_k > -x$
   $gc =< y_k > -y$
   $gd =< h_k > - < h_0 >$
4. Update weights and biases:
   $W' = W - \gamma * gW$
   $U' = U - \gamma * gU$
   $b' = b - \gamma * gb$
   $c' = c - \gamma * gc$
   $d' = d - \gamma * gd$

end for

3.2 Discriminative Training Objective

The generative training objective can be decomposed as follows:

$$L_{gen}(D_{train}) = -\sum_{i=1}^{|D_{train}|} \log p(y_i|x_i) - \sum_{i=1}^{|D_{train}|} \log p(x_i) \quad (6)$$

This means that the RBM classifier will dedicate some of its capacity at modeling the marginal distribution of the input only. Since classification is a supervised learning task and we are only interested in obtaining a good prediction of the target given the input, we can ignore the unsupervised part of the generative objective and focus on the supervised part. So, the discriminative training objective is defined as:

$$L_{disc}(D_{train}) = -\sum_{i=1}^{|D_{train}|} \log p(y_i|x_i) \quad (7)$$

The most important advantage using this training objective is that it is possible to compute exactly its

gradient with respect to the RBMs parameters for each example (x_i, y_i). For the various model parameters it stands:

- For Θ = **(c,W,U)**:

$$\frac{\vartheta \log p(y_i|x_i)}{\vartheta \Theta} = \sum_j sigm(O_{y_i j}(x_i)) \frac{\vartheta O_{y_i j}(x_i)}{\vartheta \Theta} - \sum_{j,forally} sigm(O_{yj}(x_i)) p(y|x_i) \frac{\vartheta O_{yj}(x_i)}{\vartheta \Theta} \quad (8)$$

where $O_{yj}(x) = c_j + u_{jy} + \sum_i w_{ji} x_i$ for hidden node j.

- For Θ = **(d)**:

$$\frac{\vartheta \log p(y_i|x_i)}{\vartheta d_y} = 1_{y=y_i} - p(y|x_i), \forall y \in \{1, ..., C\} \quad (9)$$

- For Θ = **(b)** the gradient is zero, since the input biases are not involved in the computation of $p(y|x)$.

3.3 Hybrid Training Objective

The effectiveness of both generative and discriminative approaches on various problems has been studied and it has been shown that they have quite different properties. For classification tasks, adding the generative training objective to the discriminative training objective is a way to regularize the second one. To adapt the amount of regularization, the Hybrid Training Objective can be used:

$$L_{hybrid}(D_{train}) = L_{disc}(D_{train}) - \alpha L_{gen}(D_{train}) \quad (10)$$

where the weight α of the generative part can be adjusted based on the performance of the model on a validation set.

3.4 Training using MNIST

In this work, RBMs were applied on a classic classification problem, handwritten digit recognition using the MNIST dataset [6]. MNIST is a large database of handwritten digits, commonly used for training various image processing systems. The database is also widely used for machine learning and pattern recognition methods. The original MNIST dataset is used here, which contains 60,000 training and 10,000 test examples with 28x28 grey-scale images corresponding to all 0-9 digits. This is a multiclass classification problem, where the number of target classes is 10. Before final integration, the best parameters for the RBM model should be selected. These parameters include the number of hidden nodes N, the learning rate γ, which training objective is minimized, the generative weight α, as well as the batch size bs. For that reason, a complete Matlab simulation model was developed and the effect of the different parameters was studied using a validation set. It should be noted that for the specific problem, Hybrid Training led to faster convergence.

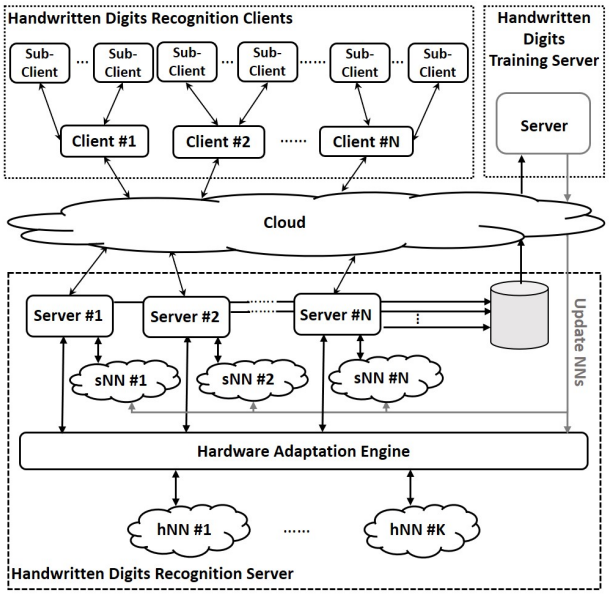


Figure 2: The HDR Computing Engine Architecture

4 HDR computing engine architecture

Based on Classification RBMs, a high performance computing engine able to serve a large amount of real-time requests for detecting the correct values of handwritten digits was implemented. This is a complete infrastructure with multiple entities specially designed either for training of already accumulated data or for real-time classification of multiple new images. Concerning the neural networks implementation, both software modules and dedicated hardware accelerators were developed.

The architecture of the proposed HDR computing engine is shown in Figure 2. This engine accepts requests from various clients from the cloud and processes them in real-time. The system is composed of three basic entities, which are a) Handwritten Digits Recognition clients, b) Handwritten Digits Recognition servers and c) Handwritten Digits (HD) training server. The HDR clients can either be individual clients or sets of sub-clients that are serviced by an HDR client that combines their requests in a single data stream. Each HDR client is associated with a dedicated HDR server. The number of HDR servers that are executed at any given time is variable and is determined by the number of HDR clients that are supported. The main advantage of the proposed architecture is that it does not use a set of predefined parameters but it uses the information of a large number of clients for continuously updating the weights and biases of the HDR algorithm, achieving

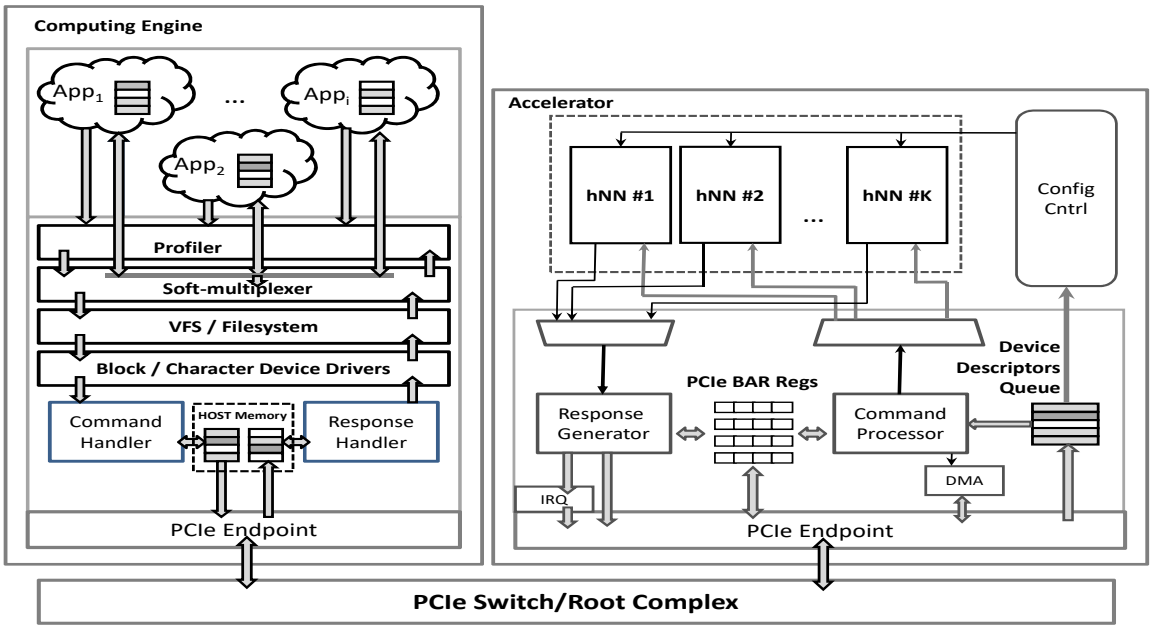


Figure 3: Hardware Adaption Engine - Interfacing with the HDR hardware accelerator

high accuracy for a given complexity. The HD training server is responsible for updating these parameters. More specifically:

HDR Clients: A two-way TCP/IP connection [19] between every HDR client and a dedicated HDR server is established. They exchange variable size data blocks, serviced either partially for better latency or as a whole block for better processing rate.

HDR Servers: When a request for a new connection arrives, a new HDR server thread is activated. Each thread receives bunches of images by its client and stores them temporarily to its local memory. Each HDR server may accept different types of requests from its client. The requests may differ at the number of images per block, the image size, i.e. 16x16 or 28x28 pixels, and pixel information coding (1 bit per pixel, 8 bits per pixel etc.). The server uses either its built-in processing capability to serve the requests, which means that the neural network is implemented in high-performance software, or forwards them to a dedicated hardware accelerator.

HD Training Server: The HD training server receives data and performs the training of the RBM neural networks. The data are forwarded to the HD training server either during initialization of the system or periodically during normal operation. In the latter case, the HDR servers send a random subset of images along with the information regarding the digit recognition. After training, the HD training server is responsible for sending the updated parameters, weights and offsets, back to all neural networks.

Concerning the neural network implementation there are three possible configurations.

(i) *Software-only Configuration:* a software NN (sNN) is initialized per HDR thread. Each NN accepts images by its corresponding thread, classifies them and returns the predicted digits.

(ii) *Hardware-only Configuration:* a small number of NNs implemented in hardware (hNN) is available to all server threads and is shared among them. The execution time is much smaller compared to the software implementation and that results to much higher processing rate. The hardware accelerator is based on a powerful FPGA board attached to the computing engine's CPU using a high-speed interface, either native PCIe Gen 3.0 with more than 1 GBps useful transfer rate [20] or the Coherent Accelerator Processor Interface (CAPI) [21]. An entity called Hardware Adaptation Engine (HAE) provides a seamless interface between the server threads and the accelerator. This entity is described analytically in the next section. Each HDR server does not use a specific hNN but whichever is available. The hardware accelerator may consist of multiple FPGA boards and/or GPUs. Although this architecture is significantly more efficient, in the case of a heavy workload, each hNN may have to process a huge bunch of images. In this case, a hybrid configuration is most preferable.

(iii) *Hybrid Configuration:* the images are still processed by the hardware accelerator (hNNs) but when the processing delay exceeds the software execution time, then the sNNs start receiving some of the images for classification.

5 Hardware Adaptation Engine

The HDR servers receive bunches of images from their clients and store them in a FIFO at the input

of HAE. Each HDR server has its own dedicated FIFO. HAE is responsible for forwarding these images for recognition to the hardware accelerator. HAE consists of multiple functional units that are shown in Figure 3. Initially, the Profiler decides the next block of images that has to be forwarded to the Soft-multiplexer. The number of images per block is selected so that the total system performance is maximized. The Soft-multiplexer interfaces with the hardware accelerator through Block/Character device drivers. The drivers and the accelerator communicate through shared host memory areas. When the images have been processed by the accelerator, a response with the digits values is fed back to the proper HDR server. A more detailed description of the HAE functional units follows.

5.1 Profiler

It is responsible for selecting the next block of images that has to be handled by the Soft-multiplexer. The Profiler decisions depend on the number of pending commands and the total response statistics (execution time etc.). The HDR servers inform the Profiler about their pending commands, while the Soft-multiplexer provides timing information regarding command execution. Initially, the Profiler selects the commands of the next block using either a round-robin or a static algorithm with fixed priorities. When enough statistics have been collected, dynamic allocation that takes into consideration the current load is feasible.

5.2 Soft-multiplexer

It makes the basic calls for interfacing with the Character/Block device drivers (e.g. open, close, pwrite, pread, ioctl etc). The Soft-multiplexer processes the selected block, extracts the commands and forwards them to the drivers. After receiving a notification for the completion of a command, the Soft-multiplexer forwards a response to the corresponding server thread. The soft-multiplexer also informs the Profiler about the execution time of each bunch of images.

5.3 Character/Block Device Driver

It is responsible for transferring the data from user space to kernel space and vice versa, for address translation and interrupt handling. The commands are associated with a data structure, in the form of a descriptor, which is the basic information provided to the hardware accelerator. The Device Driver activates the Command and Response Handler for forwarding commands and receiving the respective responses from the accelerator. The Command Handler accepts commands, creates the descriptors and stores them in a shared host memory area (Host Descriptors Queue). Each descriptor may contain one or more commands, associated with data from the same or different HDR servers. When a block of descriptors has been processed and their responses are ready, the accelerator sends an MSI-X Interrupt to notify the Response Handler that responses are stored in another shared host memory area (Responses Queue). The Response Handler decodes each response and forwards the corresponding results to the Soft-multiplexer.

At the accelerator's side, the Command Processor is responsible for accessing the descriptors at the host's main memory and transferring them into accelerator's local memory (Device Descriptors Queue). It decodes the descriptors, transfers the requested images using a DMA engine and feeds any available hNN. When the image processing has been completed the Response Generator prepares the responses that contain the classified digits, stores them in the Responses Queue and sends an interrupt to the Device Driver.

6 Neural Network Implementation

For better exploitation of the hardware resources, images of 16x16 pixels, with 1 bit/pixel, are used. This can be achieved by proper prefiltering and scaling, without affecting significantly the total system's accuracy. For this image size each hNN uses less hardware resources, thus enabling more hNNs to be integrated in the given accelerator resources, while the accuracy is only slightly reduced. These images are generated by the original ones after proper filtering and optimum threshold application. The pre-processing takes places at the client side before transmission. Before moving to the actual neural network implementation, a theoretical analysis of certain parameters regarding the prediction procedure is necessary.

6.1 Theoretical Analysis

As aforementioned, classification is based on the calculation of *free energy* [22], given by (4). For implementation purposes this equation is rearranged as follows:

$$S_m = \sum_{i=1}^{N_v} (v_i \cdot WV_{i,m}) + HB_m \tag{11}$$

$$LE_{k,m} = \log(1 + \exp(S_m + WT_{k,m})) \tag{12}$$

$$Fe_k = -(\sum_{m=1}^{N_h} LE_{k,m} + TB_k) \tag{13}$$

$$t_k = 1 : min(Fe_k) \tag{14}$$

where N_h is the number of hidden nodes, N_t the number of target nodes, $m \in [1, N_h]$, $k \in [1, N_t]$, v_i is visible node i, h_m is hidden node m and t_k is target node k. Fe_k is the *free energy* of target node k. Regarding the weights and biases, VB_i are the biases of visible nodes, TB_k are the biases of target

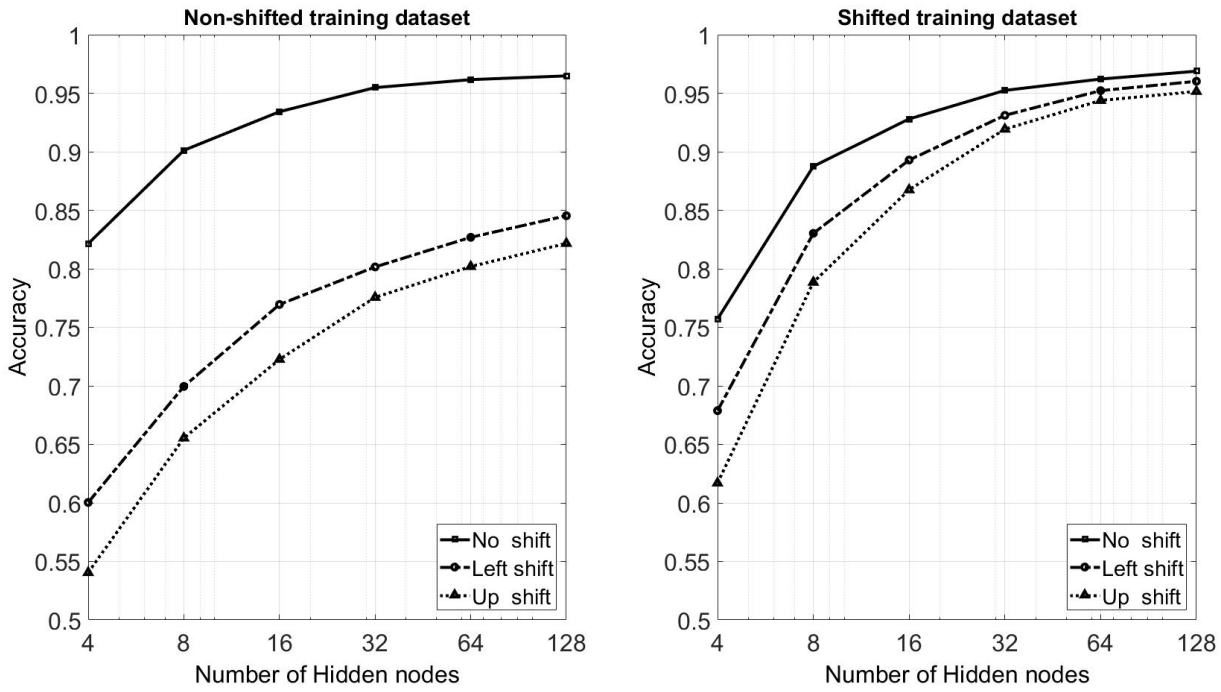


Figure 4: The influence of different training datasets on classification' s accuracy.

nodes, HB_m are the biases of hidden nodes, WV are the weights between visible and hidden nodes and WT are the weights between target and hidden nodes.

In order to study more thoroughly the accuracy of the proposed implementation and validate its effectiveness under a broader set of images, the initial dataset was extended with shifted versions of the original images. More specifically, four additional datasets were created, each one being a shifted copy of the initial one. The images were shifted to the left, right, up or down by a single pixel. In this way, the final dataset is a superset of the original one. Training was performed using both datasets, the original non-shifted one and, to keep the dimensions the same, a randomly selected collection of 60,000 images from the shifted dataset.

Figure 4 shows the effect of the training dataset on the classification accuracy of the model, when it is applied on three individual datasets, the original test dataset with the non-shifted images, the test dataset with the images that have been shifted left by one pixel and the test dataset with the images that have been shifted up by one pixel. As expected, the use of the shifted version for training leads to better overall accuracy, except in the case of the original dataset and for a small number of hidden nodes. This is explained by the fact that an RBM can model fewer dependencies when it includes a small number of hidden nodes, thus being subject to overfitting on the training dataset. Similar results can be taken for the datasets with the one-pixel right and down shifting.

Figure 4 also helps to specify the number of hidden nodes that will be used in both the software and hardware implementations of the neural network. It is obvious that the choice of 32 hidden nodes is a satisfactory trade-off between accuracy and complexity. By choosing 64 hidden nodes the accuracy improves only 2.2%, while the hardware complexity increases almost 8 times.

6.2 Software only Configuration

As described in Section 4, each HDR server receives requests from multiple clients. In software implementation each HDR server implements one RBM neural network. In order to have multiple sNNs operating in parallel, the server application is multithreaded, one thread per neural network. When a high performance CPU engine is used, multithreading results to better exploitation of the available cores. Software multithreading is mostly effective on multiprocessor or multicore systems, where actual parallel or distributed processing is feasible.

Each thread is responsible for classifying the incoming images, by calculating their *free energy* (4) with respect to each one of the ten possible classes. Therefore, vector and matrix operations dominate the computations. To achieve high performance, the threads use off-the-shelf highly-optimized libraries that exist for a variety of computer architectures.

6.3 Hardware only Configuration

Regarding the implementation of the hardware accelerator, a major consideration involves the arithmetic, fixed or floating-point, that will be

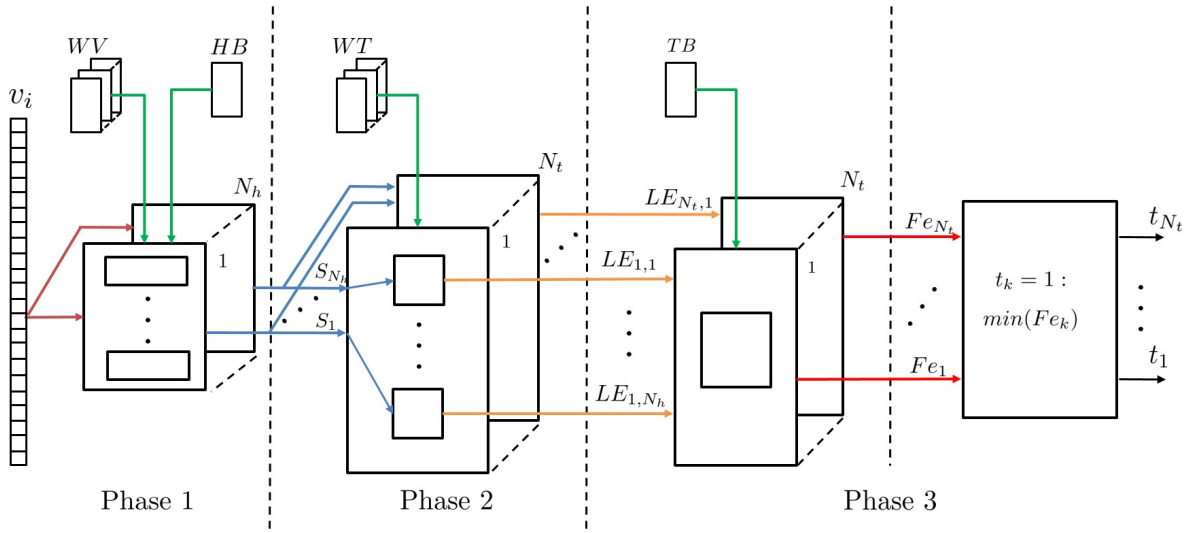


Figure 5: Accelerator Architecture

used for the calculations. The proposed design uses fixed-point arithmetic for all linear functions (*add*, *mul*, *cmp*) and single-precision floating-point for non-linear functions, like *log* and *exp* in (12). To specify the range of the fixed-point arithmetic, simulations were run and statistics were collected based on all available training and test data patterns.

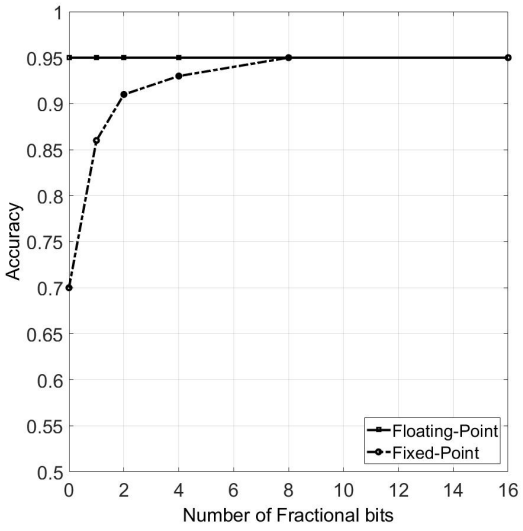


Figure 6: Accuracy vs Fractional bits ($N_h = 32$)

Figure 6 shows the accuracy results for various fixed-point number configurations. It can be seen that by using 16-bits fixed-point numbers, with the first 8 bits being used for the sign bit and the integer part and the remaining 8 bits for the fraction, a good classification accuracy of 95% can be achieved.

The proposed architecture is based on equations (11), (12) and (14) and is organized into three separate phases, shown in Figure 5. The time required to execute each phase determines the performance increase that can be achieved by proper pipeline. Whenever needed, the results of each phase are stored in shared memories and/or cascaded registers. The weights and biases are stored in dual-port RAMs and/or FIFOs, so that they are initialized/updated during system operation when new training data have been used.

A detailed description of the three phases comprising the accelerator architecture follows:

(i) *Phase 1:* Each image consists of a set of pixels, which are represented as binary values. This simplifies the multiplication part of (11), since instead of arithmetic multiplications, low complexity multiplexing functions can be used. The incoming image determines the weights that have to be used in the sum of products. Figure 7 shows a detailed scheduling of the operations implemented in this phase. Each *load* operation refers to reading WV values from memory. The *sel* of each multiplexer is connected to the corresponding pixel of the input image. This module needs 2 clocks to perform *load* and *mux* operations in order to feed the adder trees. After this point all additions are completed in 6 clocks. A total of 4 such modules is used.

Each of these modules includes 256 adders and is responsible for calculating 8 S_m results. Equation (11) determines that a total of 32 S_m results is needed. Every module operates in a pipeline manner. Although each S_m needs 8 clocks to be calculated, by using pipeline a total latency of 37 clocks is achieved.

(ii) *Phase 2:* Apart from the first addition, phase 2 is implemented using floating-point arithmetic. It was designed using the functionalities of Xilinx's Floating-point IP core [23]. The architecture of phase 2 and the respective timing diagram is given in Figure 8. Equation (12) is implemented in two similar and with same latency stages: $exp(a + b)$ and $ln(a + b)$.

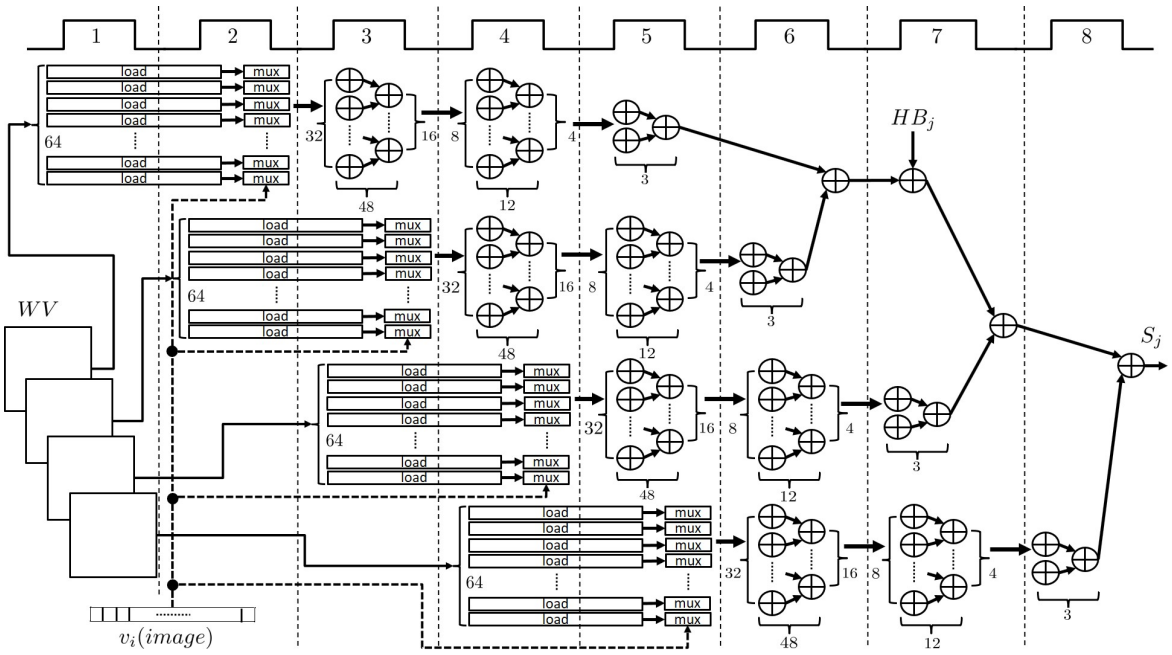


Figure 7: Phase 1 of Accelerator Architecture: Addition and Multiplexing Scheduling

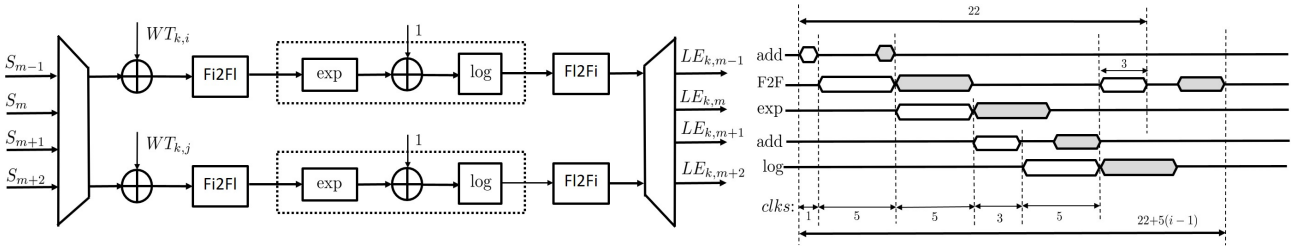


Figure 8: Phase 2 of Accelerator Architecture: Block Design (left) and Timing Diagram (right)

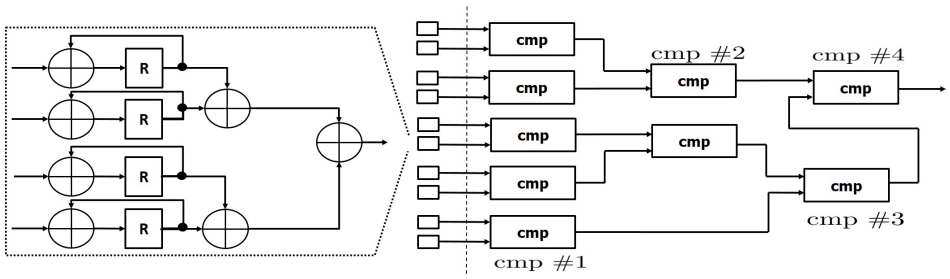


Figure 9: Phase 3 of Accelerator Architecture: Block Design

Therefore, two LE values can be calculated simultaneously using pipeline. For optimum performance, multiplexing of the incoming S_m values is performed, thus reducing the hardware complexity without decreasing the processing rate, and then fixed-to-floating point transformation is applied. Since the Fixed-point addition with the Fixed-to-Floating ($F2F$) operation lasts for less than the duration of each of the aforementioned processing stages, the same floating-point circuit (indicated with a dotted line) can be used for processing continuously a stream of S_m values. Since a total of $N_h x N_t$ LE values have to be calculated, the number of such circuits is determined by the multiplexing/demultiplexing used. $Fl2Fi$ in this figure corresponds to floating-to-fixed transformation. Phase 2 is the slowest phase of the whole accelerator. Reducing its latency would lead to improvement of the whole processing rate of the accelerator. To achieve this, sets of four LE values are calculated simultaneously. However, the required DSP resources are doubled.

(iii) *Phase 3:* The last phase is divided into two distinct sub-phases. The first accumulates the $LE_{k,j}$ inputs from Phase 2 to produce the Fe_k

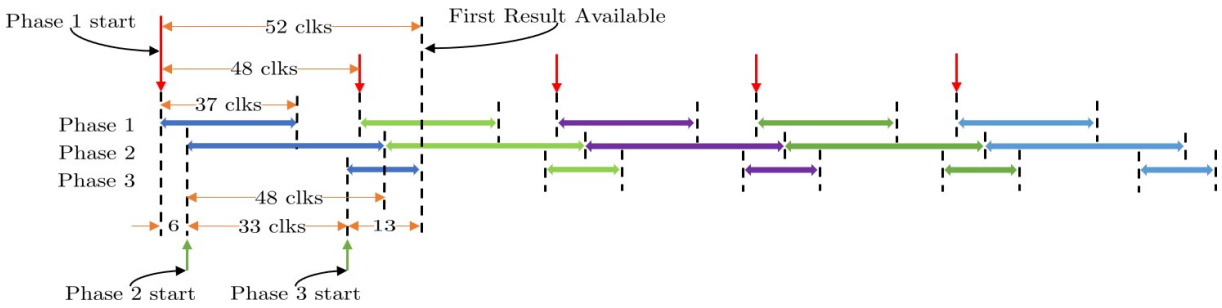


Figure 10: Pipeline strategy applied between processing phases

results (13). The second sub-phase uses the N_t computed Fe values in parallel and in a few clock cycles determines the digit (14) that corresponds to the class of the input image. Figure 9 presents the architecture of Phase 3.

Since all these phases have different processing times and do not require all the results of the previous phase to be available in order to proceed, pipeline can be used for increasing the whole system's processing rate. The latency of phase 1 is 37 clocks. In phase 2 the latency is 48 clocks and 10 such circuits are used in parallel. The combined latency for phase 3 is 10 clocks and 10 circuits implementing the accumulation process are used. Therefore the total latency per hNN without pipeline is 95 clock cycles. A pipeline of two is used and so the module is ready to receive new images every 48 clock cycles, with a latency of 52 clock cycles, as shown in Figure 10.

7 System Prototyping and Performance Results

The above described HDR Cloud Service has been implemented and tested in Xeon and Power8 servers (Table 1) while the hardware accelerator has been implemented in a Virtex-7 FPGA. For generating various workloads, an i7 server has been used, where various clients were implemented with user-defined workload patterns. The HDR clients send bunches of images over 1 Gbps ethernet link.

7.1 Software only Configuration

The applications of clients and servers were developed in C. Provided that HDR servers are multi-threaded, a very popular API is used for threading an application, known as POSIX Thread [24]. Also, CBLAS library is used for performing all the necessary vector and matrix operations. The OS of all platforms is Linux ubuntu 15.10.

In order to validate the efficiency of this configuration, the data rate achieved (in KImages/secs) for various computing servers (Table 1) and for various bunches of images (KImages/req) was measured. To preserve consistency with the hardware implementation, the number of hidden nodes used in sNNs is 32.

Table 1: Clients and Server Platforms

Platfrom	CPU		Memory		
	Cores	GHz	GB	Type	MHz
Xeon	6	1.60	16	DDR3	2133
Power8	4	3.00	64	DDR4	1600/1333
Power8	8	3.32	64	DDR4	1600/1333

When the number of images per request is small the total performance drops because the time between consecutive requests is much higher compared to the processing time of each request. On the other hand, the number of images per request does not affect the performance when it is more than a few hundreds since the overhead is absorbed. This can be seen in Tables 2 and 3.

Table 2: Processing rate [KImgs/sec] using software HDR implementation (N_t =32 and 1KImgs/req)

	Number of HDR Clients			
Servers	1	4	16	64
Intel Xeon	13.20	55.01	82.31	83.10
Power8	16.12	50.52	95.01	98.14
Power8	14.39	46.92	149.03	171.86

Table 3: Processing rate [KImgs/sec] using software HDR implementation (N_t =32 and 10KImgs/req)

	Number of HDR Clients			
Servers	1	4	16	64
Intel Xeon	13.31	55.36	82.11	82.77
Power8	16.20	59.82	98.13	100.09
Power8	14.48	54.88	155.33	171.20

Based on Tables 2 and 3, it is evident that the system performance has a linear relation with the number of HDR clients. For a small number of clients, the system is not fully utilized due to the communication overhead. Each client waits for the completion of a request before sending a new one. As the number of clients increases, this overhead is amortized and full system utilization is achieved. So, the performance increases with the increase in the number of clients. The performance reaches different maximum values, depending on the server platform

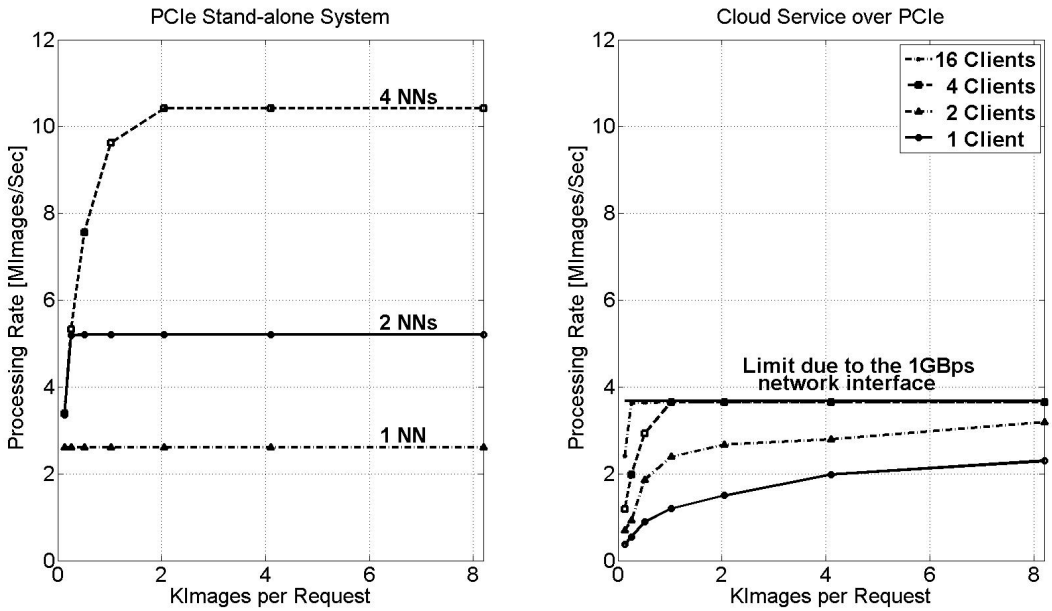


Figure 11: Processing rate using hardware HDR implementation over PCIe

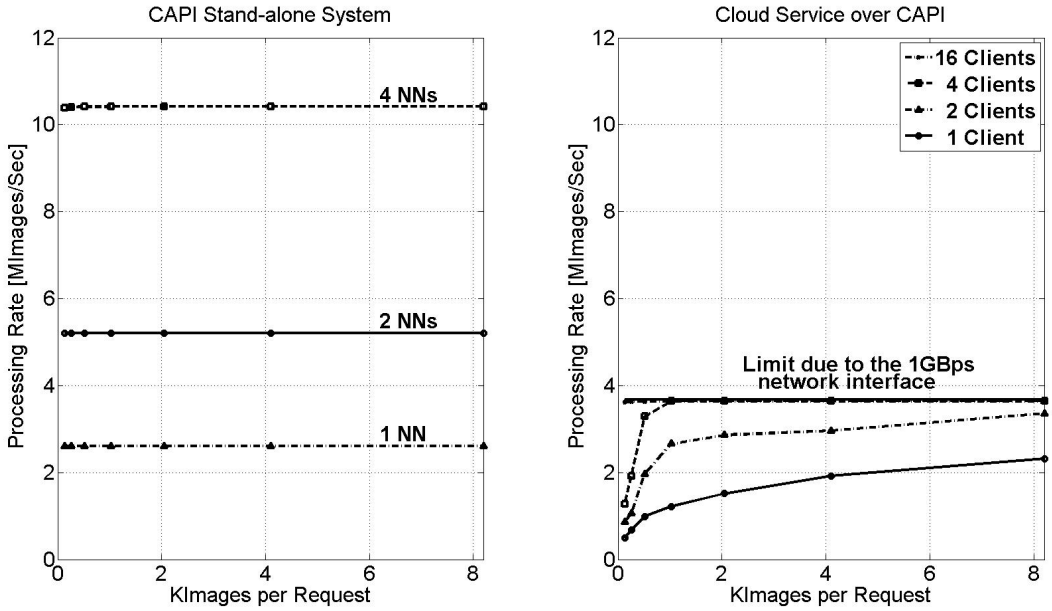


Figure 12: Processing rate using hardware HDR implementation over CAPI

used. This is mainly due to the fact that each server has different number of CPU cores (Table 1). As mentioned, when multiple threads are running, they are distributed to different cores and exploit better the capabilities of each CPU.

7.2 Hardware only Configuration

As far as the accelerator is concerned, based on the latency numbers provided above, a processing rate of 2.6 MImages/sec, at 125 MHz, is achieved by each hNN module. To maximize the total accelerator performance, four hNN modules operate in parallel. That results to a total processing rate of 10.4 MImages/sec, at 125 MHz. The limitation of four modules is introduced by the available resources of the specific Virtex-7 FPGA board (Table 4).

The accelerator supports two interfaces, PCIe Gen 3.0 with 8 lanes and Coherent Accelerator Processor Interface (CAPI). In both cases, a databus of 1024 bits is used that provides images up to four hNNs simultaneously.

The accelerator was attached to a Power8 server with 8 cores (Table 1). For each interface (PCIe, CAPI) two different sets of experiments were conducted in order to measure the performance of the hardware configuration. The results concerning the PCIe are shown in Figure 11. The left part illustrates the

accelerator's processing rate by varying the number of images per request as well as the number of hNNs implemented in the accelerator. In this case, the images are stored locally to the server and not received from multiple clients over the network. The number of images per request does not affect the performance when 1 or 2 hNNs are implemented in the accelerator. The rate is 2.6 MImages/sec and 5.2 MImages/sec correspondingly. When 4 hNNs are integrated, the maximum rate (10.4 MImages/sec) is achieved when each request contains at least 2K images.

Table 4: Implementation parameters of a neural network (N_t =32, XC7VX690T-2 Virtex-7 FPGA)

	Phase 1	Phase 2	Phases 3-4	Total
BRAMs	4 %	1 %	1 %	6 %
DSP	-	13 %	-	13 %
FFs	1 %	3 %	1 %	5 %
LUTs	7 %	8 %	1 %	16 %
Slices	-	-	-	18 %

The right part of Figure 11 illustrates the processing rate, when 4 hNNs operate in parallel, varying the number of connected clients over the network. It is obvious that the maximum rate of this set-up can be achieved even for a small number of active clients. Although in the case of 4 hNNs the maximum achievable rate is 10.4 MImages/sec it can be seen that the maximum processing rate of the whole set-up is less due to the used network interface (934.4 Mbps data rate over a 1 Gbps Ethernet link), since the communication interface becomes the systems bottleneck. Nevertheless, the system with 1 Gbps network interface and hardware acceleration is up to 21 times faster compared to the software implementation.

The achieved processing rates when CAPI is used are shown in Figure 12. The performance is slightly better when 4 hNNs are implemented in the accelerator as it can be seen in the left part of the figure. In this case, the accelerator reaches its maximum value even for a small number of images per request. So, the configuration with CAPI outperforms the configuration with native PCIe. Traditional I/O attachment protocols, like PCIe, introduce significant device driver and operating system latencies, since an application calls the device driver to access the accelerator and the device driver performs a memory mapping operation. With CAPI instead, the accelerator is attached as a coherent CPU peer over the I/O physical interface.

8 Conclusions

The design and implementation of a computing engine for handwritten digits recognition was presented. This engine can be used for cloud applications and achieves high performance, in terms of processing rate, when a hardware accelerator with multiple neural networks is used, as demonstrated by experimental results. Details of neural networks implementation on reprogrammable logic have also been described. The architecture can be parameterized in order to achieve the best compromise between hardware complexity and processing performance.

References

[1] A. Smola and S. VishwanathanKevin, *Introduction to Machine Learning*. Cambridge, UK: Cambridge University Press, 2008.

[2] A. Fischer and C. Igel, "An introduction to restricted boltzmann machines.," in *CIARP* (L. Alvarez, M. Mejail, L. G. Deniz, and J. C. Jacobo, eds.), vol. 7441 of *Lecture Notes in Computer Science*, pp. 14–36, Springer, 2012.

[3] A. Fischer and C. Igel, "Training restricted boltzmann machines," *Pattern Recogn.*, vol. 47, pp. 25–39, Jan. 2014.

[4] E. Bougioukou, N. Toulgaridis, and T. Antonakopoulos, "Cloud services using hardware accelerators: The case of handwritten digits recognition," in *Proceedings of the 6th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, (Thessaloniki), 4-6 May 2017.

[5] N. Toulgaridis, E. Bougioukou, and T. Antonakopoulos, "Architecture and implementation of a restricted boltzmann machine for handwritten digit recognition," in *Proceedings of the 6th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, (Thessaloniki), 4-6 May 2017.

[6] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," 2010. `http://yann.lecun.com/exdb/mnist/`.

[7] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pp. 1237–1242, 2011.

[8] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012*, pp. 3642–3649, 2012.

[9] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," *ArXiv e-prints*, Sept. 2015.

[10] L. Wan, M. Zeiler, S. Zhang, Y. LeCun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, 2013.

[11] R. Benenson, "Classification datasets results." `https://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html`.

[12] H.-j. Zhang and N.-f. Xiao, "Parallel implementation of multilayered neural networks based on map-reduce on cloud computing clusters," 02 2015.

[13] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," 09 2017.

[14] D. L. Ly and P. Chow, "A high-performance fpga architecture for restricted boltzmann machines," in *Proceedings of the ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, FPGA '09, (New York, NY, USA), pp. 73–82, ACM, 2009.

[15] S. K. Kim, L. C. Mcafee, P. L. Mcmahon, and K. Olukotun, "A highly scalable restricted boltzmann machine fpga implementation," in *Proceedings of the International Conference on Field Programmable Logic and Applications*, pp. 367–372, 08 2009.

[16] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, (Helsinki, Finland), pp. 536–543, 2008.

[17] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted boltzmann machine," *J. Mach. Learn. Res.*, vol. 13, pp. 643–669, Mar. 2012.

[18] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, p. 2002, 2000.

[19] M. J. Donahoo and K. L. Calvert, *TCP/IP Sockets in C: Practical Guide for Programmers*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2nd ed., 2009.

[20] PCI SIG, *PCI Express Base Specification, Revision 2.1*, March 2009.

[21] IBM Systems Magazine, *A Deeper Look at POWER8 CAPI and Data Engine for NoSQL*, May 2015. http://ibmsystemsmag.com/power/businessstrategy/competitiveadvantage/capi-deep-look/.

[22] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," in *Predicting Structured Data* (G. Bakir, T. Hofman, B. Schölkopf, A. Smola, and B. Taskar, eds.), MIT Press, 2006.

[23] Xilinx Inc., *Floating-Point Operator*. https://www.xilinx.com/products/intellectual-property/floating_pt.html.

[24] B. Nichols, D. Buttlar, and J. P. Farrell, *Pthreads Programming*. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 1996.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 1, 496-501(2018)

www.astesj.com

Special issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

An Analysis of K-means Algorithm Based Network Intrusion Detection System

Yi Yi Aung*, Myat Myat Min

Faculty of Computer Sciences, University of Computer Studies, Mandalay, UCSM, 0000, Myanmar

ARTICLE INFO

Article history:
Received: 30 November, 2017
Accepted: 30 January, 2018
Online: 10 February, 2018

Keywords:
Network Intrusion Detection System
K-means
Random Forest
KDDCup 99

ABSTRACT

In this modern age, information technology (IT) plays a role in a number of different fields. And therefore, the role of security is very important to control and assist the flow of activities over the network. Intrusion detection (ID) is a kind of security management system for computers and networks. There are many approaches and methods used in ID. Each approach has merits and demerits. Therefore this paper highlights the similar distribution of attacks nature by using K-means and also the effective accuracy of Random Forest algorithm in detecting intrusions. This paper describes full pattern recognition and machine learning algorithm performance for the four attack categories, such as Denial-of-Service (DoS) attacks (deny legitimate request to a system), Probing attacks (information gathering attacks), user-to-root (U2R) attacks (unauthorized access to local super-user), and remote-to-local (R2L) attacks (unauthorized local access from a remote machine) shown in the KDD 99 Cup intrusion detection dataset.

1. Introduction

On the Internet, users share valuable information around the world. The internet has created various ways to threaten the stability and security of interrelated systems. Both of these mechanisms are static and dynamic. Static mechanisms like firewalls and software updates provide dynamic security and mechanisms such as intrusion detection systems. Today, security is the most serious problem for getting valuable information. Therefore, static mechanisms or dynamic mechanisms are needed to protect individual information despite the precautionary technology. The intrusion detection system detects not only successful aggression, but also helps monitor and prevent timely action.

The intrusion Detection System (IDS) is a standard component of a security infrastructure that allows network administrators to detect policy violations. Check all incoming and outgoing network activity and determine suspicious patterns that indicate network or system attacks from people trying to break or compromise the system.

A secure network must provide the following:

- Data confidentiality: Data transferred over the network must be accessible only to data that has been approved accordingly.
- Data integrity: Data must maintain integrity from when it is sent when it is received. No damage or loss of data from random events or malicious activities is accepted.
- Data availability: The network must be resistant to service attack denial.

IDS technology based on tracking process can be categorized into two approaches:

Abuse/Signature detection: This technology searches for signature attacks and known signatures in network traffic and are used as a reference to detect future attacks. Regularly updated databases are usually used to store signatures of known attacks. The way this technology controls intrusion detection is similar to antivirus software. The advantage of this type of detection is that it can accurately and efficiently detect known attacks. Anomaly detection: This technology is based on tracking traffic anomalies. The gap between traffic is monitored and regular profiles are measured. Different implementations of this technology have been reserved based on metrics used to measure the deviation of traffic profiles. The advantage of this detection type is that it is well suited to detect unknown attacks.

IDS are divided into two parts based on analysis and retention of audit data:

* Yi Yi Aung, Email: yiyiaungresearch@gmail.com

Host-based IDS (HIDS): HIDS is a home based tracking method that allows the system to collect data in the form of multiple host activity records, such as event logs and system logs. Since everything is in the host, there's no need to install additional hardware or software [1]. The advantages of hosted IDS are to check the success or failure of attacks, monitor system activity, and detect attacks that IDS networks cannot detect, close tracking and real-time responses, are not required.

Network-based IDS (NIDS): NIDS is a network approach that collects data directly from a network monitored as a packet instead of collecting data from a particular host / agent. Most NIDS are a free and easy-to-use operating system [2]. Network-based IDS offers advantages such as low cost of ownership, easier, placement, network attack detection, evidence preservation, real-time tracking and rapid response, and detection of failed attacks.

2. Literature Review

Most of intrusion detection system focused on four major attack categories such as denial of service, probe, user-to-root, and remote-to-local but this author specially emphasized on User-to-Root (U2R) attacks in NSL-KDD dataset by using Weka tool. This paper focused on a comparative study analysis of user-to-root attack, which the attacker tries to access normal user account and gains root access information of the system based on several machine learning techniques such as navie bayes, random forest, J48, etc [3].

This paper analyzed anomaly intrusions detection system by using Random Forest classifier with Principal Component Analysis. The author got experimental results by using simulation connection dataset of NSL-KDD. The performance of the system was measured by using Precision, Recall and F-Measure. And also this paper was specially focused on to detect various attacks present in Denial of Service (DoS) such as Neptune, Smurf, Pod, Teardrop, Land, Back, Apache2, Processtable, Mailbomb [4].

This paper used C4.5, CART (Classification and Regression Trees), Random Forest, and REP (Reduced Error Pruning) Tree to investigate the detection of intrusions contained in KDDCUP 1999 DARPA dataset. And compared the performance of the above algorithms based on the measures such as Accuracy, Learning Time (in seconds) and Size of the Tree. According to the experimental results, Random Forest was better as it correctly identifies more number of instances than other. And the accuracy of the REP Tree was very less than other algorithms but the learning time of REP Tress is very less than other [5].

They used Support Vector Machine with Principal Component Analysis (PCA) to choose the optimum feature subset that was useful in applying for intrusion detection system. To determine the effectiveness and feasibility of the proposed IDS system, they choosed NSL-KDD dataset for simulation their system. They found that PCA algorithm is good to select a best subset of features for classification of intrusions. It can help to speed up the training and testing process of intrusions detection which is important for high-speed network applications [6].

In this proposed paper, several classification techniques and machine learning algorithms have been considered to categorize the network traffic. Out of the classification techniques, they have found nine suitable classifiers like BayesNet, J48, PART, JRip, Random Tree, Random Forest and REPTree. The comparison of these algorithms has been performed using WEKA tool [7].

Security has become a crucial issue for computer systems. IDS can protect to our computer network. Different classification and clustering algorithms have been proposed in recent year for IDS. In this paper, multiple algorithms were analyzed to find the optimal algorithm. At last the optimal algorithms Random Forest and DB Scan were occurred for IDS [8].

The purpose of this survey paper was to describe the methods/ techniques which are being used for Intrusion Detection based on Data mining concepts and the designed frame works. This survey paper stated the methods and techniques of data mining to aid the process of Intrusion Detection and the frameworks [9]. The concept of intercepting these two different fields, gives more scope for the research community to work in this area. New approaches enhanced the existing interference detecting system and it was a stepping stone to build effective and efficient IDS to detect different types of attacks [10].

This paper proposed a novel hybrid model for intrusion detection. The proposed framework in this paper may be expected as another step towards advancement of IDS. The Hybrid framework led to effective, adaptive and intelligent intrusion detection [11].

This paper drew the conclusions on the basis of implementations performed using various data mining algorithms. Combining more than one data mining algorithms had be used to remove disadvantages of one another and lead to a better performance than any single classifier. Different classifiers had different knowledge regarding the problem [12].

3. Methodology

This section consists of the conversation of the two algorithms of data mining classification approaches. These are K-means and Random Forest.

3.1. K-means Clustering Algorithm

Clustering, based on distance measurements performed on objects, and classifying objects (invasions) into clusters. Unlike classification, classification because there is no information about the label of learning data is an unattended learning process. For anomalous detection, we can use welding and in-depth analysis to guide the ID model. Measurement of distance or similarity plays an important role in collecting observations into homogeneous groups. Jacquard affinity measurement, the longest common order scale (LCS), is important that the event is to awaken the size to determine if normal or abnormal. Euclidean distance is approximately two vectors X and Y in space Euclidean n-dimensions, the size of the distance widely used for vector space. Euclidean distance can be defined as the square root of the total difference of the same vector dimension. Finally, grouping and classification algorithms need to be channeled effectively, massively, it possible to handle dimension of network data and heterogeneity [13].

In this paper, we use K-means algorithm to cluster dataset connections. The K-means algorithm is one of the widely recognized clustering tools. K-means groups the data in accordance with their characteristic values into a user-specified number of K distinct clusters. Data categorized into the same cluster have identical feature values. K, the positive integer denoting the number of clusters, needs to be provided in advance. The steps involved in a K-means algorithm are given consequently: [14]

1. K points denoting the data to be clustered are placed into the space. These points denote the primary group centroids.
2. The data are assigned to the group that is adjacent to the centroid.
3. The positions of all the K centroids are recalculated as soon as all the data are assigned.
4. Repeat steps 2 and 3 until the centroid unchanged.

This results in the partition of data into groups. The preprocessed dataset partition is performed using the K-means algorithm with K value as 5. Because we have the dataset that contains normal and 4 attack categories such as DoS, Probe, U2R, R2L.

3.2. Random Forest Algorithm

One of the most popular methods or frameworks used by scientists in the science of data is Random Forest. It is a supervised classification algorithm. It can be seen from its name, which is to create a forest by some way and make it random. There is a direct relationship between the numbers of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach [15].

Random Forests grows many classification trees. Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random – but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number mM is specified such that at each node, m variables are selected at random out of the M and the best split on this m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

There are many of top benefits of Random Forest algorithm. Some of these benefits are as follows:

- Accuracy
- Runs efficiently on large data bases
- Handles thousands of input variables without variable deletion
- Provides effective methods for estimating missing data
- Maintains accuracy when a large proportion of the data are missing

4. KDDCup 99 Dataset

The evaluation of any intrusion detection algorithm on real network data is extremely difficult mainly due to the high cost of obtaining proper labeling of network connections. Due to the real sample cannot be gotten for intrusion detection, the KDDCup'99 dataset is used as the sample to verity the performance of the misuse detection model. The KDDCup'99 dataset, referred by Columbia University, was arranged from intrusions simulated in a military network environment at the DARPA in 1998. It contains network connections obtained from a sniffer that records all network traffic using the TCP dump format. The total simulated period is seven weeks. It was performed in the MIT Lincoln Labs, and then announced on the UCI KDD Cup 1999 Archive [16].

KDDCup'99 dataset have two variations of training dataset; one is a full training set having 5 million connections and the other is 10% of this training set having 494021 connections. Since the whole dataset is huge, the experiment has been performed on its smaller amount of dataset that is 10% of KDD. Additionally, the KDDCup'99 dataset includes many attack behaviors, classified into four groups: Probe, Denial of Service (DoS), User to Root (U2R), and Remote to Local (R2L) [17]. These can be seen in table I. Normal connections are created to profile that expected in a military network. The detailed information of the two variations of training dataset can be seen in table II.

Table I: Various Attacks and Categories

Categories	Attacks Subclass
DoS	back, land, Neptune, pod, smurf, teardrop
Probe	ipsweep, nmap, portsweep, satan
U2R	buffer_overflow, loadmodule, perl, rootkit
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster

Table II: Number of Instances in KDD and 10% KDD

Class	Whole KDD	10 % KDD
DoS	3883370	391458
Probe	41102	4107
U2R	52	52
R2L	1126	1126
Normal	972780	97278
Total	4898430	494021

The data set includes 41 features classifying the data records into normal or a type of attacks. The features consist of 34 types of numeric features and 7 types of symbolic features, according to different properties of attack. The nature of features can be divided into the following groups [18].

- Basic Features: Basic functions can be obtained from the packet header without checking the load.

- Content Features: Domain knowledge is used to assess the original TCP packet load. This includes features such as the number of unsuccessful login attempts.
- Time-based Traffic Features: This function is designed to capture properties in the 2-second window. Examples of such functions are the number of connections to the same host every two seconds.
- Host-based Traffic Features: Use the history window to estimate the number of connections (in the case 100) and not the time. Therefore, host-based functionality is designed to assess attacks that include two or more intervals.

4.1. Pre-Processing

KDDCUP 99 data set is pre-processed in order to make it suitable for the data mining learning algorithm. Pre-processing is performed for the following reasons.

Each record in the dataset consists of categorical as well as numeric features. Textual (plain) data is used for categorical features. K-means algorithm needs numeric data (either discrete or continuous). The first step in pre-processing is to covert this categorical feature attributes to numeric attributes. For converting symbols into numerical form, an integer code is assigned to each symbol. For instance, in the case of protocol type feature, 0 is assigned to tcp, 1 to udp, and 2 to the icmp symbol and so on. The dataset contains three categorical attributes while the rest of the thirty eight attributes are numeric. Every category of an attribute is assigned a specific number.

We have used K-means and Random Forest to define normal and attacks in the system. They need specific format so we have converted the dataset to K-means and Random Forest compatible format.

5. Experimental Results and Discussion

To facilitate the experiments, we used eclipse java and weka tool to implement the algorithms on a PC with 64-bit window 7 operating system, 4GB RAM and a CPU of Intel core i3-4010U CPU with 1.70GHz. Data come from MIT Lincoln laboratory of KDDCup99 data set. The table lists the number of instances available in the whole dataset, 10% of KDDCup'99 dataset.

The analysis is performed by using K-means and Random Forest algorithms. We use K-means algorithm to generate heterogeneous dataset to nearly homogeneous dataset. The clustering results of K-means algorithm are described from table III to table VIII.

Table III: Detailed Information of Attack Categories in Cluster-1

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	107219	107217	2
Probe	1610	1605	5
U2R	0	0	0
R2L	6	3	3
Normal	10	3	7
Total	108845	108828	17

Table IV: Detailed Information of Attack Categories in Cluster-2

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	1067	1065	2
Probe	1221	1207	14
U2R	4	0	4
R2L	1	0	1
Normal	21235	21230	5
Total	23528	23502	26

Table V: Detailed Information of Attack Categories in Cluster-3

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	280782	280782	0
Probe	0	0	0
U2R	0	0	0
R2L	0	0	0
Normal	16	14	2
Total	280798	280796	2

Table VI: Detailed Information of Attack Categories in Cluster-4

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	2203	2202	1
Probe	12	1	11
U2R	46	29	17
R2L	1087	1068	19
Normal	75409	75398	11
Total	78757	78698	59

Table VII: Detailed Information of Attack Categories in Cluster-5

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	187	186	1
Probe	1264	1255	9
U2R	2	0	2
R2L	32	22	10
Normal	608	604	4
Total	2093	2067	26

Table VIII: Detailed Information of Attack Categories with Clustering

Attacks	Total Records	Correctly Classify Records	Incorrectly Classify Records
DoS	391458	391452	6
Probe	4107	4068	39
U2R	52	29	23
R2L	1126	1093	33
Normal	97278	97249	29
Total	494021	493891	130

By analyzing the clustering results, the characteristics of Denial of Service (DoS) attacks are mostly related to themselves in cluster-3. And then, it is closely similar to the nature of Probe attacks in cluster-1. Probe attacks are also mostly related to DoS attacks in cluster-1. And then, it is nearly same with the nature of

Normal by looking in cluster-5. Normal is mostly similar nature with User-to-Root attacks and Remote-to-Local attacks by studying in cluster-4. And then, Normal is related to Probe by studying cluster-2 and cluster-5. Normal is related to all attacks by looking in all 5 clusters because attacks mimic to normal behavior in intrusions.

Then we apply Random Forest algorithm to know the intrusions and normal traffic. The performance of attacks categories with Random Forest algorithm in 5 clusters of K-means can be seen from table IX to table XIV. The Precision and Recall of the normal and attacks detection are good and the false positive rate is nearly zero.

Table IX: Performance Analysis of Attack Categories in Cluster-1

Attacks	False Positive Rate	Precision	Recall
DoS	0.00738	0.999888	0.999981
Probe	0.000009	0.999377	0.996894
U2R	0	0	0
R2L	0.000018	0.6	0.5
Normal	0.000018	0.6	0.3

Table X: Performance Analysis of Attack Categories in Cluster-2

Attacks	False Positive Rate	Precision	Recall
DoS	0	1	0.998125
Probe	0.000224	0.995874	0.988533
U2R	0.000042	0	0
R2L	0	0	0
Normal	0.008722	0.999058	0.999764

Table XI: Performance Analysis of Attack Categories in Cluster-3

Attacks	False Positive Rate	Precision	Recall
DoS	0.125	0.999992	1
Probe	0	0	0
U2R	0	0	0
R2L	0	0	0
Normal	0.875	0	1

Table XII: Performance Analysis of Attack Categories in Cluster-4

Attacks	False Positive Rate	Precision	Recall
DoS	0	1	0.999546
Probe	0	0	0.083333
U2R	0.000165	0.690476	0.630434
R2L	0.000077	0.994413	0.98252
Normal	0.000495	0.999483	0.999854

Table XIII: Performance Analysis of Attack Categories in Cluster-5

Attacks	False Positive Rate	Precision	Recall
DoS	0.001049	0.989361	0.994652
Probe	0.00965	0.993665	0.992879
U2R	0.000478	0	0
R2L	0.002911	0.785714	0.6875
Normal	0.00606	0.985318	0.993421

Table XIV: Performance Analysis of Attack Categories with K-means Clustering

Attacks	False Positive Rate	Precision	Recall
DoS	0.000156	0.999959	0.999984
Probe	0.000028	0.99657	0.990504
U2R	0.00003	0.65909	0.557692
R2L	0.000028	0.987353	0.969831
Normal	0.000148	0.99928	0.999701

6. Conclusion

This paper presents a comparative analysis hybrid machine learning technique to detect Denial of Service (DoS) attacks, Probing (Probe) attacks, User-to-Root (U2R) attacks and Remote-to-Local (R2L) attacks. We can know the similar nature of attack group by using K-means algorithm. And then we use Random Forest algorithm to classify normal and attack connections. The experiments show that, KDDCup 99 dataset can be applied as an effective benchmark dataset to help researchers compare different intrusion detection models. Future work includes analyzing with other data mining algorithms to classify attack categories and how it can detect on other real time environment dataset.

References

[1] Aung Yi Yi and Myat Myat Min, "An Analysis of Random Forest Algorithm Based Network Intrusion Detection System", Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNDP), 2017 IEEE/ACIS 18 th International Conference, 2017.

[2] Yan. K.Q., S. C. Wang and C. W. Liu, "A Hybrid Intrusion Detection System of Cluster-based Wireless Sensor Networks", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, March 18-20, 2009, Hong Kong.

[3] S. Revathi and Dr. A. Malathi, "Detecting User-To-Root (U2R) Attacks Based on Various Machine Learning Techniques", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 4, ISSN (Online): 2278-1021, ISSN (Print): 2319-5940, April 2014.

[4] S. Revathi and Dr. A. Malathi, "Detecting Denial of Service Attack Using Principal Component Analysis with Random Forest Classifier", International Journal of Computer Science & Engineering Technology (IJCSET), Vol.5, No. 03, ISSN: 2229-3345, March 2014.

[5] Jayshri R. Patel, "Performance Evaluation of Decision Tree Classifiers for Ranked Features of Intrusion Detection", Journal of Information, Knowledge and Research in Information Technology, Vol. 02, Issue -02, ISSN: 0975-6698, Nov 12 to Oct 13.

[6] Heba F. Eid et al., "Principal Components Analysis and Support Vector Machine based Intrusion Detection System", 10 th International Conference on Intelligent Systems Design and Applications, (IEEE, 2010).

[7] Manzoor, Muhammad Asif, and Yasser Morgan, "Network Intrusion Detection System using Apache Storm" , Special Issue on Recent Advances in Engineering Systems, Advances in Science, Technology and Engineering System Journal (ASTES), Vol. 2, No. 3, 812-818 (2017).

[8] S. Choudhury and A. Bhowal, "Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection", International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015, pp 89-95.

[9] R. Venkatesan, R. Ganesan and A.A.L. Selvakumar, "A Comprehensive Study in Data Mining Frameworks for Intrusion Detection", International Journal of Advanced Computer Research, December-2012, Volume-2 Number-4 Issue-7, ISSN (print): 2249-7277 ISSN (online): 2277-7970.

[10] Somani Manish and Roshni Dubey, "Hybrid Intrusion Detection Model Based on Clustering and Association", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.3, Issue 3, ISSN (Print): 2320-3765, ISSN(Online):2278-8875 March 2014.

[11] M. Dhakar and A. Tiwari, "A Novel Data Mining based Hybrid Intrusion Detection Framework", Journal of Information and Computing Science, 2014, Vol-9 No-1 pp. 037-048, ISSN 1746-7659, England, UK.

[12] TR. Patel, A. Thakkar and A. Ganatra, "A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems", International Journal of Soft Computing and Engineering (IUSCE), March-2012, Vol-2, Issue-1, ISSN: 2231-2307.

[13] Youssef Ahmed and Ahmed Emam, "Network Intrusion Detection Using Data Mining and Network Behavior Analysis", International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011.

[14] X. Wu, V.Kumar, Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg, "Top 10 algorithms in data mining", Survey Paper(2008).

[15] S. Devaraju and S. Ramakrishnam, "Performance Comparison for Intrusion Detection System using Neural Network with KDD Dataset", ICTACT Journal on soft Computing , Vol:04, Issue:03, ISSN: 2229-6956, April 2014.

[16] P. S. Rath, M. Hohanty, S. Acharya and M. Aich, "Optimization of IDS Algorithms Using Data Mining Technique", Proceeding of 53rd IRF International Conference, Pune, India,ISBN 978-93-86083-01-2, 2016.

[17] L.S. Parihar and A. Tiwari, "Survey on Intrusion Detection Usingn Data Mining Methods",IJSART, January-2016, Volume-2 Issue-1 ISSN (online): 2395-1052.

[18] Md.E. Haque and T.M. Alkharobi, "Adaptive Hybrid Model for Network Intrusion Detection and Comparison among Machine Learning Algorithms", International Journal of Machine Learning and Computing, February 2015, Vol-5, No-1.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 502-510 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Computation of Viability Kernels on Grid Computers for Aircraft Control in Windshear

Nikolai Botkin[*,1], Varvara Turova[1], Johannes Diepolder[2], Florian Holzapfel[2]

[1]*Zentrum Mathematik, Modeling M6, Technische Universität München, 85748, Germany*

[2]*Institute of Flight System Dynamics, Technische Universität München, 85748, Germany*

ARTICLEINFO

Article history:
Received: 29 November, 2017
Accepted: 24 January, 2018
Online: 28 February, 2018

Keywords:
Aircraft Cruise Flight Phase
Differential Game
Viability Kernel
Grid Method
Supercomputer

ABSTRACT

This paper is devoted to the analysis of aircraft dynamics in the cruise flight phase under windshear conditions. The study is conducted with reference to a point-mass aircraft model restricted to move in a vertical plane. We formulate the problem as a differential game against the wind disturbances: The first player, autopilot, manages, via additional smoothing filters, the aircraft's angle of attack and power setting. The second player, wind, produces disturbances that are transferred, also via smoothing filters, into most dangerous wind gusts. The state variables of the game are subject to state constraints representing aircraft safety conditions related, for example, to the altitude, path inclination and velocity. Viability theory is used to compute the so-called viability kernels, the maximal subsets of state constraints where an appropriate feedback strategy of the first player can keep aircraft trajectories arbitrary long for all admissible disturbances generated by the second player. A grid method is utilized, and challenging computations in seven dimensions are conducted on a supercomputer system.

1 Introduction

Atmospheric conditions such as windshear continue to be considered as a source of potentially severe consequences. They are dangerous for aircraft during landing or take-off, because the wind gusts can occur at relatively low altitudes. Nevertheless, windshear is also dangerous during the cruise flight phase because it can lead to violation of the prescribed flight level.

In view of threats related to wind disturbances, there is permanent interest in designing robust aircraft guidance and control schemes (possibly for use with autopilots). The related question consists in finding safety domains, i.e. sets of initial states from which the control problem can be solved in the case of worst wind disturbance whose components lie in a known range.

There exist a large number of works devoted to the problem of aircraft control in the presence of severe windshears. In particular, papers [1–6] address the problem of aircraft control during take-off in windshear conditions. In works [1] and [2], the wind velocity field is assumed to be known. It is shown that open loop controls obtained as solutions of appropriate optimization problems provide satisfactory results for rather severe wind disturbances. Nevertheless, it is clear that the spatial distribution of wind velocity cannot be measured with appropriate accuracy, and therefore feedback principles of control design are more realistic. Different types of feedback controls are proposed in papers [3–6]. In [3], the design of a feedback robust control is based on the construction of an appropriate Lyapunov function. Robust control theory is used in [4] to develop feedback controls stabilizing the relative path inclination and (in [5] and [6]) for the design of feedback controls stabilizing the climb rate.

An approach based on differential game theory (see e.g. [7]) is presented in paper [8] in connection with the problem of landing. A high dimensional nonlinear system of dynamic equations is linearized and reduced to a two-dimensional differential game using a transformation of variables. The resulting differential game is numerically solved, and optimal feedback controls are constructed and tested in the nonlinear model against a downburst.

[*]Corresponding Author: Nikolai Botkin, Boltzmannstr. 3, 85748 Garching bei München, Germany, Email: botkin@ma.tum.de

https://dx.doi.org/10.25046/aj030161

Another method, also based on differential game theory, is used in papers [9] and [10] to find feedback controls that are effective against downbursts. This approach assumes the computation of the value function, which is a viscosity solution (see. e.g. [11] and [12]) of an appropriate Hamilton-Jacobi equation. The numerical implementation utilizes dynamic programming techniques described in [13] and [14]. The case of known wind velocity field as well as the case of unknown wind disturbances are considered.

Recent investigations [15] and [16] utilize a method (see [17]) for the fast computation of rough approximations of solvability sets in linear conflict control problems. Using techniques of sequential linearization, this method is applied to nonlinear aircraft dynamics to design an appropriate control for take-off in the presence of downbursts.

The current paper addresses the problem of retaining trajectories in an appropriate flight domain (AFD) corresponding to the cruise flight phase (cf. [18] and [19]). Viability theory (see e.g. [20]) provides numerical methods (see e.g. [21], [22], and the Appendix) for finding the viability kernel, i.e. the largest set of initial states lying in the AFD from which viable trajectories emanate. More precisely, it includes all initial states for which there exists a feedback control that generates trajectories remaining in the viability kernel for all possible admissible wind gusts. In the case where the initial state does not belong to the viability kernel, there exists a method of designing a wind disturbance such that all trajectories violate the AFD for all possible controls.

As for wind conditions, it is assumed that only bounds on the wind velocity components are imposed. The dynamics of the aircraft will be considered as a differential game (cf. [9] and [10]) where the first player chooses control inputs, whereas the second player forms the worst wind disturbance. It is assumed that the first player is able to measure the current state vector, whereas the second player can measure both the current state vector and the current control ("future" values are not available) of the first player. Therefore, the second player may use the so-called feedback counter strategies (see [7]).

The current paper has common features with the open-access publication [23] concerning the model description and solution method. It should be noted that the publication [23] is mainly focused on theoretical fundamentals of the differential game approach. Regarding computational results, paper [23] formulates a problem of constructing the viability kernel in seven dimensions and performs several steps of the algorithm to show its feasibility. In contrast, the current paper addresses aspects of implementation on large scale grid computers and completely solves the above mentioned seven-dimensional problem including simulation of optimal trajectories.

The paper is structured as follows:

Section 2 outlines a point-mass aircraft model describing the vertical motion of a generic modern regional jet transport aircraft. The model is closely related to the one described in paper [23]. The difference consists in a more clear method of deriving the dynamics equations.

In Section 3, state constraints related to the cruise flight phase are formulated, and the corresponding computed viability kernels are demonstrated through their three-dimensional sections. Additionally, trajectories yielded by an optimal feedback strategy, working against an optimal control of the disturbance, are shown. It should be noted that an optimal control of the disturbance can be constructed either as counter or pure feedback strategy because the so called saddle point condition holds for the differential game under consideration.

Section 4 outlines some aspects of parallel implementation of the computational method on a supercomputer system. The parallelization principles and data flow inside and between compute nodes are sketched. The novelty of our approach and comparison with existing software tools are discussed.

Section 6 (Appendix) briefly outlines the concept of differential games and viability kernels. Grid schemes for the computation of them are sketched. The details can be found in [23].

2 Model equations

In this section, a point-mass model representing the vertical motion of a generic modern regional jet transport aircraft is considered. Table 1 introduces euclidean coordinate systems (COS) that are necessary to compute the forces exerted on the aircraft. The origins of COSs are located either at the aircraft gravity center (CG) or at a fixed reference point (O) on the earth surface.

Table 1: Aircraft coordinate systems

COS	**Index**	***x*-axis**	**Origin**
Local	N	Parallel to the Earth's surface (x_N) and upwards (z_N)	O
Kinematic	K	In direction of $\overrightarrow{V_K}$	CG
Aerodynam.	A	In direction of $\overrightarrow{V_A}$	CG
Thrust	P	In positive direction of the symmetry axis of the turbine (aft looking forward)	CG
Body Fixed	B	In direction of the nose and in the symmetry plane of the aircraft	CG

Here, $\vec{V}_K$ and $\vec{V}_A$ are kinematic and aerodynamic aircraft velocities, respectively. The angles defining the relationship between the coordinate systems are the following (see also Figure 1):

α_K the kinematic angle of attack,
α_A the aerodynamic angle of attack,
γ_K the kinematic path inclination angle,
σ the thrust inclination angle.

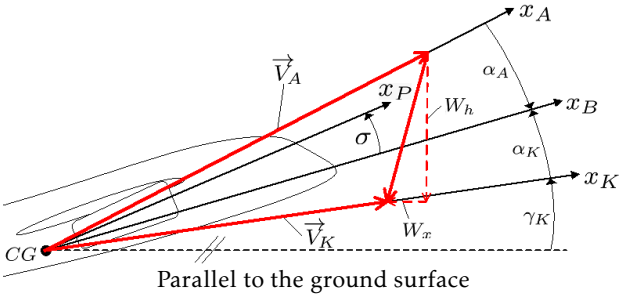


Figure 1: Aircraft coordinate systems and angles

Matrices for the transformations $K \to N$ (Kinematic to Local), $A \to B$ (Aerodynamic to Body), $B \to K$ (Body to Kinematic), and $P \to B$ (Thrust to Body) are defined as follows:

$$M_{NK} = \begin{bmatrix} \cos(\gamma_K) & -\sin(\gamma_K) \\ \sin(\gamma_K) & \cos(\gamma_K) \end{bmatrix},$$

$$M_{BA} = \begin{bmatrix} \cos(\alpha_A) & -\sin(\alpha_A) \\ \sin(\alpha_A) & \cos(\alpha_A) \end{bmatrix},$$

$$M_{KB} = \begin{bmatrix} \cos(\alpha_K) & -\sin(\alpha_K) \\ \sin(\alpha_K) & \cos(\alpha_K) \end{bmatrix},$$

$$M_{BP} = \begin{bmatrix} \cos(\sigma) & -\sin(\sigma) \\ \sin(\sigma) & \cos(\sigma) \end{bmatrix}.$$

The position propagation is given in the local coordinate system (N), whereas the translation dynamics are derived in the kinematic coordinate system (K). The model equations read as follows:

$$\dot{x}_N = V_K \cos(\gamma_K), \quad (1)$$
$$\dot{z}_N = V_K \sin(\gamma_K), \quad (2)$$
$$\dot{V}_K = \frac{X_T}{m}, \quad (3)$$
$$\dot{\gamma}_K = \frac{Z_T}{m V_K}. \quad (4)$$

Here, X_T and Z_T denote the components of the total force $\vec{F}_T$ represented in the kinematic coordinate system (K), and the notation m stands for the aircraft mass. As usually, $\vec{F}_T$ comprises aerodynamic, propulsion, and gravitation forces:

$$\vec{F}_T = \vec{F}_A + \vec{F}_P + \vec{F}_G.$$

Aerodynamic forces. They are defined as follows:

$$\vec{F}_A = M_{KB} M_{BA} \begin{bmatrix} C_D \\ C_L \end{bmatrix} \frac{1}{2} \rho V_A^2 S,$$

where C_D and C_L are the drag and lift coefficients, respectively, $\rho = \rho(h)$ is the air density (depends on the altitude), V_A the aerodynamic velocity, and S the wing reference area.

The lift and drag coefficients $C_D(\alpha_A, M)$ and $C_L(\alpha_A, M)$ are taken in the form:

$$C_D(\alpha_A, M) = c_1^D + c_2^D \alpha_A + c_3^D M + c_4^D \alpha_A^2 + c_5^D \alpha_A M + c_6^D M^2 + c_7^D \alpha_A^3 + c_8^D \alpha_A^2 M + c_9^D \alpha_A M^2, \quad (5)$$

$$C_L(\alpha_A, M) = c_1^L + c_2^L \alpha_A + c_3^L M + c_4^L \alpha_A^2 + c_5^L \alpha_A M + c_6^L M^2 + c_7^L \alpha_A^3 + c_8^L \alpha_A^2 M + c_9^L \alpha_A M^2, \quad (6)$$

where the coefficients c_i^D and c_i^L, $i = 1,...9$ are determined from least square fitting to experimental data.

The absolute value, V_A, of the aerodynamic velocity can be derived using its relation to the kinematic velocity $\vec{V}_K$ in the Local frame (N) and the wind velocities W_x and W_z in the x_N and z_N directions, respectively. Therefore,

$$V_A = \left\| (\vec{V}_K)^N - \begin{bmatrix} W_x \\ W_z \end{bmatrix} \right\|, \quad (7)$$

and finally, using the matrix M_{NK}, this implies the formula

$$V_A^2 = (V_K \cos\gamma_K - W_x)^2 + (V_K \sin\gamma_K - W_z)^2.$$

The aerodynamic angle of attack α_A is computed as follows:

$$\alpha_A = \arctan\left(\frac{w_A}{u_A}\right),$$

where u_A and w_A are x_B and z_B-components of the aerodynamic velocity, respectively.

The Mach number M is defined as follows:

$$M = \frac{V_A}{c}, \quad c = \sqrt{\kappa R T(h)},$$

where c is the speed of sound, κ the adiabatic index for air, R the gas constant for ideal gases, and $T(h)$ the temperature of air at the altitude h. See [23] for more details.

Propulsion Forces. Thrust forces are modeled considering a two-engine setup. Thus,

$$\vec{F}_P = 2 M_{KB} M_{BP} \vec{F}_{P\text{net}}, \quad \vec{F}_{P\text{net}} = \begin{bmatrix} f_V(\delta_T, M) \\ f_\gamma(\delta_T, M) \end{bmatrix},$$

where $\delta_T \in [0,1]$ is the thrust setting, and the functions f_V and f_γ are approximated similar to formulas (5) and (6), with δ_T instead of α_A. See [23] for more details.

Gravitation Force. For a simple gravitational model with constant acceleration g, the corresponding force is computed as:

$$F_G = M_{KN} \begin{bmatrix} 0 \\ -g \end{bmatrix},$$

where M_{KN} is the inverse (transpose) of M_{NK}.

The following two equations are added to the dynamics (1)-(4) to exclude jumps in the controls:

$$\dot{\alpha}_K = \widetilde{\alpha}_K, \quad \dot{\delta}_T = \widetilde{\delta}_T \tag{8}$$

Furthermore, the following equations produce smoothing of wind disturbances:

$$\dot{W}_x = -k_w(W_x - \widetilde{W}_x), \quad \dot{W}_z = -k_w(W_z - \widetilde{W}_z), \tag{9}$$

where the time constant, k_w, is chosen as $k_w = 1\,s^{-1}$.

3 Problem setting and simulation results

Problem. The model consists of equations (2), (3), (4), (8), and (9). Thus, the state vector has seven variables: z_N, V_K, γ_K, α_K, δ_T, W_x, and W_z. The controls are associated with the rate of the angle of attack, $\widetilde{\alpha}_K$, and the rate of the thrust setting, $\widetilde{\delta}_T$. Their instantaneous changes are permitted. The disturbances are now associated with the artificial variables $\widetilde{W}_x$ and $\widetilde{W}_z$ that are inputs of the filters (9). Thus, the physical wind components W_x and W_z do not exhibit instantaneous jumps.

The following constraints on the controls and disturbances are prescribed:

$$\begin{gathered}\widetilde{\alpha}_K \in [-5,5]\,\text{deg/s}, \;\; \widetilde{\delta}_T \in [-0.3,0.3]\,1/\text{s}, \\ |\widetilde{W}_x| \le 5\,\text{m/s}, \;\; |\widetilde{W}_z| \le 5\,\text{m/s},\end{gathered} \tag{10}$$

and the following state constraints are imposed:

$$\begin{aligned}&h_N := z_N - h_0 \in [-90,90]m, \\ &V_K \in [100,200]\,\text{m/s}, \;\; |\gamma_K| \le 10\,\text{deg}, \\ &\alpha_K \in [0,16]\,\text{deg}, \;\; \delta_T \in [0.3,1],\end{aligned} \tag{11}$$

where $h_0 = 10000\,\text{m}$ being the cruise flight altitude.

Additionally, the state constraints $|W_x| \le 5\,\text{m/s}$ and $|W_z| \le 5\,\text{m/s}$ hold automatically because of equations (9) and the constraints (10).

In the numerical construction, the box $[-100,100]\times[90,210]\times[-15,15]\times[-4,20]\times[0.2,1.2]\times[-6,6]\times[-6,6]$ of the space $(h_N, V_K, \gamma_K, \alpha_K, \delta_T, W_x, W_z)$ was divided in $200\times120\times30\times24\times14\times12\times12$ grid cells. The sequence of time steps, $\{\delta_\ell\}$, was chosen as $\delta_\ell \equiv 0.01$, and the computations were performed until $\left|\mathcal{V}^h_{\ell+1} - \mathcal{V}^h_\ell\right| \le 10^{-5}$ for all grid nodes. Totally, 4471 steps of the algorithm (19) were done. The computation has been done on the SuperMUC system at the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities. The computation was distributed over 100 compute nodes with 16 cores per node, which is regarded as "middle task" on the SuperMUC system. The runtime was about 40 h.

Figures 2-4 show different three-dimensional sections of the seven-dimensional viability kernel. Figures 5-7 respectively demonstrate the same three-dimensional sections and the corresponding projections of two optimal trajectories emanating from points lying in the viability kernel. The start point of trajectory 0 lies near to the boundary of the viability kernel, whereas trajectory 1 starts from a point lying deep inside of the viability kernel. The trajectories are computed when the control uses its optimal feedback strategy, and the disturbance utilizes its optimal counter feedback strategy (see the Appendix). The time step size used in the simulation of the trajectories was equal to 0.01. It is seen that the trajectories go to their attraction cycles and remain there. The simulation time interval is $[0, 15\,\text{min}]$. Figure 8 shows a three-dimensional section of a smaller viability kernel corresponding to the shrinkage of the state constraints (11) by the factor 0.7. The corresponding three-dimensional projection of two optimal trajectories are shown. It is seen that the disturbance can keep trajectory 0 outside the reduced viability kernel because the start point lies outside of it.

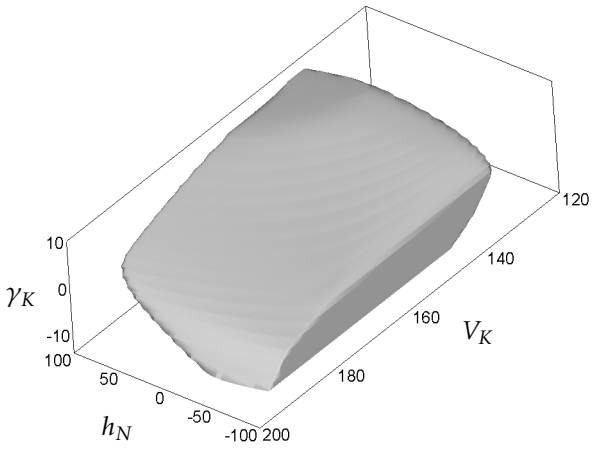


Figure 2: A three-dimensional section of the viability kernel: $\alpha_K = 8$, $\delta_T = 0.65$, $W_x = 0$, and $W_z = 0$.

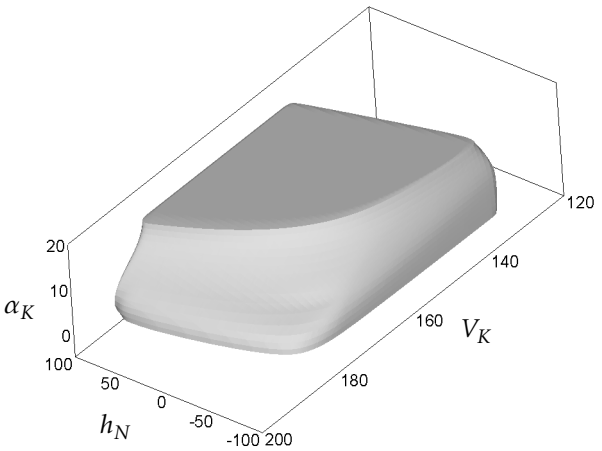


Figure 3: Another three-dimensional section of the viability kernel: $\gamma_K = 0$, $\delta_T = 0.65$, $W_x = 0$, and $W_z = 0$.

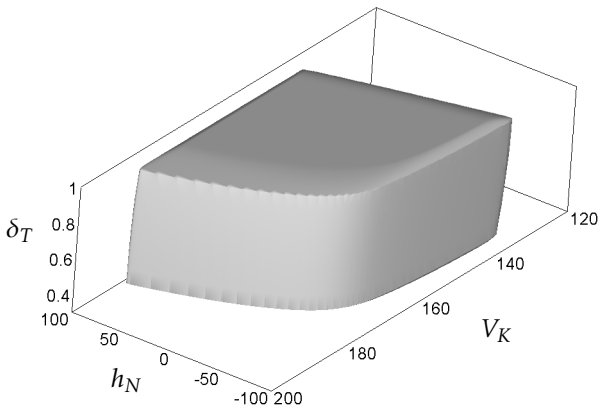


Figure 4: One more three-dimensional section of the viability kernel: $\gamma_K = 0$, $\alpha_K = 8$, $W_x = 0$, and $W_z = 0$.

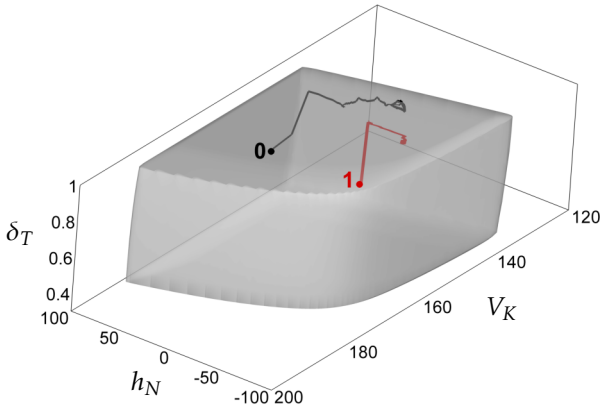


Figure 7: The same section as in Figure 4 and projections of the same two trajectories as in Figure 5.

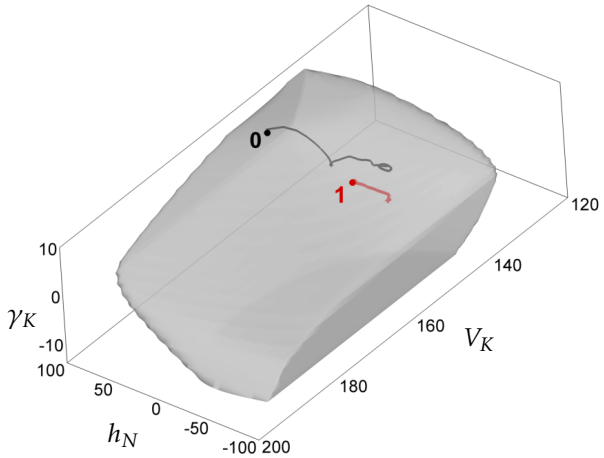


Figure 5: The section from Figure 2 and projections of two trajectories generated by an optimal feedback strategy of the control and an optimal counter feedback strategy of the disturbance. The start points are marked with bullets.

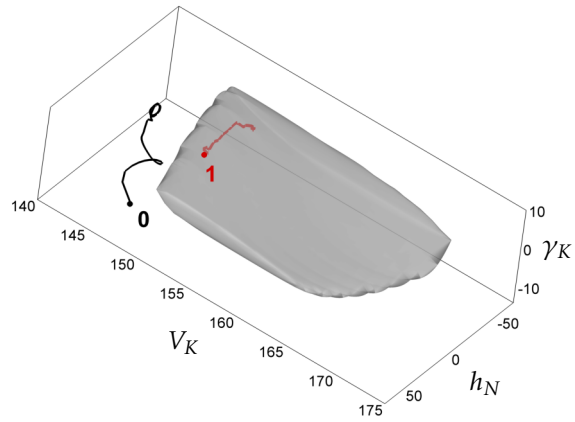


Figure 8: The section ($\alpha_K = 8$, $\delta_T = 0.65$, $W_x = 0$, and $W_z = 0$) of a smaller viability set corresponding to the shrinkage of the state constraints by the factor of 0.7. Projections of the same trajectories as in in Figure 5 are shown. Since the start point 0 does not belong to this viability set, the disturbance can keep the trajectory outside of it.

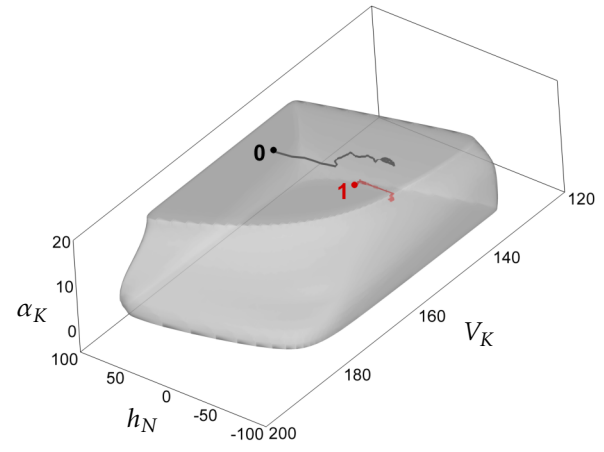


Figure 6: The same section as in Figure 3 and projections of the same two trajectories as in Figure 5.

4 Implementation Aspects

4.1 Grid Computing, Parallelization and Scaling

We use a self developed software code to implement the algorithms of the grid scheme outlined in Appendix 6.3. The code is parallelized using a mixed MPI/OpenMP technique. The first two dimensions of the grid are decomposed, whereas the other dimensions remain unmodified. A compute node cartesian topology (see Figure 9) is then created such that each compute node corresponds to a grid cylinder. Additionally, each grid cylinder is supplied with ghost nodes that allow us to compute divided differences (18) used e.g. in the algorithm (19). In Figure 9, the

lines connecting the compute nodes show the data flow supported by MPI. The parallelization inside of each compute node (i.e. inside a grid cylinder) is supported by OpenMP.

The problem considered in this paper deals with a seven-dimensional grid of 200×120×30×24×14×12×12 cells. Each of the first two dimensions was divided into 10 parts. Thus, there were 100 grid cylinders, each of size $20 \times 12 \times 30 \times 24 \times 14 \times 12 \times 12$, plus the necessary ghost grid nodes. Therefore, the required memory per compute node was equal to about 8 GB.

Our observations show a good scaling behavior (see Table 2). The results were obtained for the problem of computing the viability kernel in five dimensions on a $200 \times 120 \times 30 \times 24 \times 14$ grid. The relative speedup normalized to 32 cores and absolute timing of 30000 steps are shown.

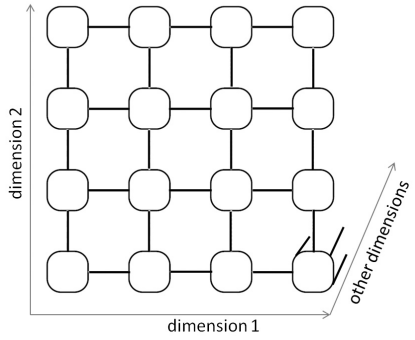


Figure 9: Compute node cartesian topology used in our applications.

Table 2: Scaling behavior of the code

# of cores	Linear prediction of scaling	Observed scaling	Wall time [h]	Performance/core [GFlop/s]
32	1	1	9.7	20
128	4	3.5	2.8	17.5
1600	50	38	0.25	15.2

4.2 Novelty of the method used

This paper utilizes a new method for computing viability kernels, which is based on the results of paper [21]. It is proven in [21] that the viability kernel is the Hausdorff limit of the sets $\{x : V(x,t) \le 0\}$ as $t \to -\infty$, where $V(x,t)$ is the value function (see [7]) of a state constrained differential game .

The numerical implementation of this method requires a theoretical basis and a stable numerical procedure for the treatment of transient Hamilton-Jacobi equations related to differential games with state constraints. Such a theoretical basis is given in paper [13] where the conventional conditions for viscosity solutions are modified to account for state constraints. This theoretical background is numerically implemented in [13] and [14], which results in monotone stable grid scheme (17) and (19). This enables us to perform a large number of time steps, usually several thousands.

Another new feature is that the corresponding software code is deeply parallelized using hybrid MPI/OpenMP techniques and adapted to run on a supercomputer system. Moreover, diverse modified variants of the code, based on sparse representations of grid functions, are tested.

4.3 Comparison with existing software

The well known software for solving Hamilton-Jacobi equations is the Level Set Method Toolbox (LSMT) described in [24]. This tool is really appropriate for solving rather general problems. However, it is not indicated in the manual whether the LSMT can compute viability kernels for differential games with *state constraints*. Moreover, according to the manual, the LSMT is not parallelized, whereas our software runs on a multi-node system.

5 Conclusion

This investigation shows that methods of differential games theory and viability theory can by applied to nonlinear aircraft models to investigate potential control abilities in the presence of wind disturbances. The new feature of our approach is the consideration of viability kernels for differential games. Feedback strategies of the players can be found from limiting grid "value functions" defining viability kernels (cf. Appendix). It is important that the amount of stored data is relatively small, which permits to implement the computed feedback strategies on a flight simulator. Further research will be focused on the treatment of models with more state variables, which will allow us to consider more realistic problems. Moreover, accounting for additional sources of uncertainty such as sensor errors or modeling uncertainties is planned.

Acknowledgment This work was supported by the DFG grants TU427/2-1 and HO4190/8-1. Computer resources for this project have been provided by the Gauss Centre for Supercomputing/Leibniz Supercomputing Centre under grant: pr74lu.

6 Appendix

This section briefly summarizes a method presented in [23] for computing viability kernels. Such a summary should help to provide a self-contained presentation.

6.1 Differential game

Consider a conflict control system with the autonomous dynamics

$$\dot{x} = f(x,u,v),\ x \in R^n,\ u \in P \subset R^p,\ v \in Q \subset R^q. \quad (12)$$

Here; x stands for the state vector; u and v denote control inputs of the first and second players, respectively; and the compact sets P and Q describe constraints imposed on the control and disturbance variables, respectively. Further, it is assumed that all functions of x have global properties. For example, the right-hand side f is supposed to be bounded, continuous, and Lipschitzian in x on $R^n \times P \times Q$.

In the following, it is assumed that the Isaacs saddle point condition

$$\min_{u\in P}\max_{v\in Q}\langle s, f(x,u,v)\rangle = \max_{v\in Q}\min_{u\in P}\langle s, f(x,u,v)\rangle, \quad (13)$$

is true for all $s \in R^n$ and $x \in R^n$. Note that this condition holds for the problem under consideration because controls and disturbances are additively separated in the model equations.

6.2 Viability kernel

For any $v \in Q$, consider the differential inclusion

$$\dot{x} \in F_v(x) = \overline{\text{co}}\{f : f(x,u,v), u \in P\} \quad (14)$$

Let $G \subset R^n$ be a compact set such that $G = \overline{\text{int}\, G}$, this set will play the role of the state constraint. Let T be an arbitrary fixed time instant, and $N = (-\infty, T] \times G$. For any subset $W \subset N$ and any time instant $t \le T$, define the time section of W through the relation $W(t) := \{x \in R^n : (t,x) \in W\}$.

Definition 1 (u-stability property [7]) *A set $W \subset N$ is called u−stable on $(-\infty, T]$ if for any position $(t_*, x_*) \in W$, for any time instant $t^* \in [t_*, T]$, for any fixed $v \in Q$, there exists a solution $x(\cdot)$ of the differential inclusion (14) with the initial condition $x(t_*) = x_*$ such that $(t^*, x(t^*)) \in W$.*

The next proposition is taken as a basis of the definition of viability kernels.

Proposition 1 *[see [21] for the proof] Let W be a maximal u-stable subset of $N = (-\infty, T] \times G$. If $W(t) \neq \varnothing$ for any $t \le T$, then the set*

$$K = \bigcap\nolimits_{t\le T} W(t)$$

is nonempty, and $W(t) \to K$ in the Hausdorff metric as $t \to -\infty$. The set K is called the viability kernel of G and denoted by $Viab(G)$.

Proposition 2 *Let $\bar{t} > 0$ be an arbitrary time instant, and $x_* \in Viab(G)$. Then there exists a feedback strategy $\mathcal{A}(x)$ of the first player such that all trajectories yielded by $\mathcal{A}$ and started at $t = 0$ from x_* remain in the set $Viab(G)$ for all $t \in [0, \bar{t}]$ and any actions of the second player. If $x^* \notin Viab(G)$, then there exists a feedback strategy $\mathcal{B}(x)$ and a time instant t_f such that all trajectories yielded by $\mathcal{B}$ and started at $t = 0$ from x^* violate the state constraint G for $t > t_f$ and any actions of the first player.*

6.3 Grid method for computing viability kernels

For the implementation of numerical method, it is convenient to represent viability kernels as level sets of an appropriate function. Let G_λ be a family of state constraint sets defined by the relation

$$G_\lambda = \{x \in R^n, g(x) \le \lambda\}, \quad (15)$$

where a continuous function g is chosen in such a way that, e.g., G_0 being the desired state constraint. It is necessary to construct a function V representing the viability kernels as follows:

$$Viab(G_\lambda) = \{x \in R^n, V(x) \le \lambda\}. \quad (16)$$

Such a function can be computed as a limiting solution, as $t \to -\infty$, of an appropriate Hamilton-Jacobi equation arising from conflict control problems with state constraints (see [13]). A grid approximation of V is computed as described below (cf. [21], [22], and [23]).

Let $\delta > 0$ be a time step length, and the tuple $h := (h_1, ..., h_n)$ defines space step sizes. Set $|h| := \max\{h_1, ..., h_n\}$ and introduce the following upwind operator defined on grid functions related to the discretization h:

$$\Pi(\phi;\delta,h)(x) = \phi(x) + \delta \min_{u\in P}\max_{v\in Q}\sum_{i=1}^{n}(p_i^{\text{right}} f_i^{+} + p_i^{\text{left}} f_i^{-}), \quad (17)$$

with f_i being the components of $f(x,u,v)$, and

$$\begin{aligned} &a^+ = \max\{a, 0\}, \quad a^- = \min\{a, 0\}, \\ &p_i^{\text{right}} = [\phi(x_1, ..., x_i + h_i, ..., x_n) - \phi(x_1, ..., x_i, ..., x_n)]/h_i, \\ &p_i^{\text{left}} \ = [\phi(x_1, ..., x_i, ..., x_n) - \phi(x_1, ..., x_i - h_i, ..., x_n)]/h_i. \end{aligned} \quad (18)$$

Let $\{\delta_\ell\}$ be a sequence of positive reals such that $\delta_\ell \to 0$ and $\sum_{\ell=0}^{\infty}\delta_\ell = \infty$. Consider the following grid scheme:

$$\mathcal{V}^h_{\ell+1} = \max\left\{\Pi\left(\mathcal{V}^h_\ell; \delta_\ell, h\right), g^h\right\}, \ \mathcal{V}^h_0 = g^h, \quad \ell = 0, 1, \ldots, \quad (19)$$

where g^h is the restriction of g to the grid defined by h.

It can be proven that $\mathcal{V}^h_\ell$ monotonically converges point-wise to a grid function $\mathcal{V}^h$, and this function define approximations of the viability kernels according to formula (16), see [23] for more details.

Remark 1 *In (17), the operation $\min_{u\in P}\max_{v\in Q}$ can be changed for $\max_{v\in Q}\min_{u\in P}$ to obtain almost the same result. The difference tends to zero with $|h|$. The proof follows from the fact that the original operator (17) and the modified one satisfy the same consistency condition (see [13]) involving the following Hamiltonian H:*

$$H(x,p) := \max_{v\in Q}\min_{u\in P}\langle p, f(x,u,v)\rangle = \min_{u\in P}\max_{v\in Q}\langle p, f(x,u,v)\rangle.$$

Numerical computations confirm this observation.

6.4 Control design

This section outlines one of possible methods of control design. Consider the grid scheme (19) assuming that ℓ is large enough so that the required approximation is reached, i.e. $\left|\mathcal{V}^h_{\ell+1} - \mathcal{V}^h_{\ell}\right|_{L_\infty} \le \epsilon$.

The optimal control u and the worst disturbance $v(u)$ at the current state x of the game can be found as solutions of the following program:

$$u, v \to \min_{u\in P} \max_{v\in Q} \mathcal{L}^h\Big[\mathcal{V}^h_\ell\Big]\Big(x + \tau f(x,u,v)\Big). \qquad (20)$$

Here, $\mathcal{L}^h$ is an interpolation operator (see e.g. [14]) defined on the corresponding grid functions, and τ being a parameter which should be several times larger than the time step size of the simulation procedure to provide some stabilization. Note that the function $\mathcal{V}^h_\ell$ can be transferred to a sparse grid (see e.g. [25] and [26]), which may essentially reduce the storage space. The disadvantage of such a technique is a certain loss of accuracy and a slower performance.

Remark 2 *As it is described above, the second player uses the so-called feedback counter strategies, i.e. functions of x and u, where u being the current control action of the first player. Since the saddle point condition (13) holds, the theory of differential games says that the second player can achieve the same result using pure feedback strategies, i.e. functions of x. For example, a near optimal strategy of the second player can be obtained as a solution of the problem*

$$u, v \to \max_{v\in Q} \min_{u\in P} \mathcal{L}^h\Big[\mathcal{V}^h_\ell\Big]\Big(x + \delta f(x,u,v)\Big),$$

see also Remark 1. Thus, the both players achieve optimal results using pure feedback strategies.

Remark 3 *If u and v appear linearly in the right-hand side of (12), then* min *and* max *operations in (17) and (20) can be computed only over extreme points of the sets P and Q respectively. This can be proven using the same arguments as in Remark 1.*

Remark 4 *If x_0 being a start point of the game, then x_0 lies in the approximate viability kernel defined as:*

$$\{x \in R^n : \mathcal{V}^h_\ell(x) \le \mathcal{V}^h_\ell(x_0)\},$$

and all trajectories approximately remain in this set. However, if the second player works non-optimally for a while, then, most likely, $\mathcal{V}^h_\ell(x(\bar{t})) < \mathcal{V}^h_\ell(x_0)$ for some $\bar{t} > 0$, and therefore, the state vector $x(\bar{t})$ lies now in the smaller viability kernel

$$\{x \in R^n : \mathcal{V}^h_\ell(x) \le \mathcal{V}^h_\ell(x(\bar{t}))\}.$$

Thus, faults of the second player improve the result of the first one.

Conflict of Interest The authors declare no conflict of interest.

References

[1] A. Miele, T. Wang, W. W. Melvin, "Optimal take-off trajectories in the presence of windshear" J. Optimiz. Theory App., **49**(1), 1–45, 1986. https://doi.org/10.1007/BF00939246

[2] A. Miele, T. Wang, W. W. Melvin, "Guidance strategies for near-optimum take-off performance in windshear" J. Optimiz. Theory App., **50**(1), 1–47, 1986. https://doi.org/10.1007/BF00938475

[3] Y. H. Chen, S. Pandey, "Robust control strategy for take-off performance in a windshear" Optim. Contr. Appl. Met., **10**(1), 65–79, 1989. doi:10.1002/oca.4660100106

[4] G. Leitmann, S. Pandey, "Aircraft Control under Conditions of Windshear" in C. T. Leondes (ed.) Control and Dynamic Systems, vol. 34, part 1, 1–79, 1990. Academic Press, New York.

[5] G. Leitmann, S. Pandey, "Aircraft control for flight in an uncertain environment: takeoff in windshear" J. Optimiz. Theory App., **70**(1), 25–55, 1991. https://doi.org/10.1007/BF00940503

[6] G. Leitmann, S. Pandey, E. Ryan, "Adaptive control of aircraft in windshear" Int. J. Robust Nonlin., **3**(2), 133–153, 1993. doi:10.1002/rnc.4590030206

[7] N. N. Krasovskii, A. I. Subbotin, Game-Theoretical Control Problems, Springer, 1988.

[8] V. S. Patsko, N. D. Botkin, V. M. Kein, V. L. Turova, M. A. Zarkh, "Control of an aircraft landing in windshear" J. Optimiz. Theory App., **83**(2), 237–267, 1994. https://doi.org/10.1007/BF02190056

[9] N. D. Botkin, V. L. Turova, "Application of dynamic programming approach to aircraft take-off in a windshear" AIP Conference Proceedings, **1479**, 1226–1229, 2012. https://doi.org/10.1063/1.4756373

[10] N. D. Botkin, V. L. Turova, "Dynamic programming approach to aircraft control in a windshear" in V. Křivan, G. Zaccour(eds.) Advances in Dynamic Games. Annals of the International Society of Dynamic Games, vol. 13, 53–69, 2013. Birkhäuser, Cham. https://doi.org/10.1007/978-3-319-02690-9_3

[11] M. G. Crandall, P. L. Lions, "Viscosity solutions of Hamilton-Jacobi equations" T. Am. Math. Soc., **277**(1), 1–47, 1983. https://doi.org/10.1090/S0002-9947-1983-0690039-8

[12] A. I. Subbotin, Generalized Solutions of First Order PDEs: The Dynamical Optimization Perspective, Birkhäuser, 1995.

[13] N. D. Botkin, K.-H. Hoffmann, N. Mayer, V. L. Turova, "Approximation schemes for solving disturbed control problems with non-terminal time and state constraints" Analysis, **31**(4), 355–379, 2011. https://doi.org/10.1524/anly.2011.1122

[14] N. D. Botkin, K.-H. Hoffmann, V. L. Turova, "Stable numerical schemes for solving Hamilton–Jacobi–Bellman–Isaacs equations" SIAM J. Sci. Comput., **33**(2), 992–1007, 2011. https://doi.org/10.1137/100801068

[15] K. Martynov, N. Botkin, V. Turova, J. Diepolder, "Real-Time Control of Aircraft Take-Off in Windshear. Part I: Aircraft Model and Control Schemes" in 25th Mediterranean Conference on Control and Automation (MED), Valletta Malta, 277–284, 2017. IEEE. doi:10.1109/MED.2017.7984131

[16] K. Martynov, N. Botkin, V. Turova, J. Diepolder, "Real-Time Control of Aircraft Take-Off in Windshear. Part II: Simulations and Model Enhancement" in 25th Mediterranean Conference on Control and Automation (MED), Valletta Malta, 285–290, 2017. IEEE. doi:10.1109/MED.2017.7984132

[17] E. K. Kostousova, "On target control synthesis under set-membership uncertainties using polyhedral techniques" in C. Pötzsche, C. Heuberger, B. Kaltenbacher, F. Rendl (eds.) System Modeling and Optimization, vol. 443, 170–180, 2014. Springer-Verlag, Berlin-Heidelberg. https://doi.org/10.1007/978-3-662-45504-3_16

[18] N. Seube, R. Moitie, G. Leitmann, "Aircraft take-off in windshear: a viability approach". Set-Valued Anal., **8**(1-2), 163–180, 2000. https://doi.org/10.1023/A:1008786811464

[19] A. M. Bayen, I. M. Mitchell, M. K. Osihi, C. J. Tomlin, "Aircraft autolander safety analysis through optimal control-based reach set computation" J. Guid. Control Dynam., **30**(1), 68–77, 2007. https://doi.org/10.2514/1.21562

[20] J. P. Aubin, "A survey of viability theory" SIAM J. Control Optim., **28**(4), 749–788, 1990. https://doi.org/10.1137/0328044

[21] N. D. Botkin, V. L. Turova, "Numerical construction of viable sets for autonomous conflict control systems" Mathematics, **2**(2), 68–82, 2014. doi:10.3390/math2020068

[22] N. D. Botkin, V. L. Turova, "Examples of computed viability kernels" Trudy Inst. Mat. i Mekh. UrO RAN, **21**(1), 306–319, 2015. http://mi.mathnet.ru/eng/timm1190

[23] N. Botkin, V. Turova, J. Diepolder, M. Bittner, F. Holzapfel, "Aircraft control during cruise flight in windshear conditions: viability approach" Dyn. Games Appl., **7**(4), 1–15, 2017. https://doi.org/10.1007/s13235-017-0215-9

[24] I. Mitchell, A Toolbox of Level Set Methods, UBC Department of Computer Science Technical Report TR-2007-11, 2007. https://www.cs.ubc.ca/~mitchell/ToolboxLS/toolboxLS-1.1.pdf

[25] C. Zenger, "Sparse grids" in Hackbusch, W. (ed.) Parallel Algorithms for Partial Differential Equations. Notes on Numerical Fluid Mechanics, vol. 31, 241–251, 1991. Vieweg, Braunschweig/Wiesbaden. doi: 10.1002/zamm.19920721115

[26] D. Pflüger, "Spatially Adaptive Sparse Grids for High-Dimensional Problems", Dissertation, Verlag Dr. Hut, München, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 511-520 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Adaptive and Non Adaptive LTE Fractional Frequency Reuse Mechanisms Mobility Performance

Uttara Sawant*, Robert Akl

Department of Computer Science and Engineering, University of North Texas, Denton, Texas 76207, USA

ARTICLEINFO

Article history:
Received: 30 November 2017
Accepted: 5 February 2018
Online: 28 February, 2018

Keywords:
Wireless
Metrics
Optimization
Frequency reuse
Mobility
Femtocells

ABSTRACT

Mobile broadband has gained momentum with the growing demand of user data rates. Long Term Evolution (LTE) technology is the step in mobile communications evolution, developed to satisfy high data rate demand, and meet better spectral efficiency requirements. Effective radio resource management and inter cell interferences are the major challenges. Fractional Frequency Reuse (FFR) is one of the effective interference avoidance mechanisms applied to LTE networks to yield optimal throughput. In this extended paper, we propose a novel performance metric, weighted throughput on user satisfaction, and evaluate an existing adaptation process that dynamically adjusts to optimal network performance determined by FFR mechanism with mobile users. The performance of FFR mechanism with mobility model, adaptation process, and femtocell densification is evaluated and optimized for proposed metric and other metrics. Results optimized by proposed metric show comparatively higher average throughput and lower variance among user throughput.

1 Introduction

Orthogonal Frequency Multiple Access (OFDMA) offers great spectrum efficiency and flexible frequency allocation to users. However, in LTE OFDMA networks the system performance is severely hampered by the Inter-Cell Interference (ICI) due to the frequency reuse, where the cell edge users will experience high interferences from neighboring cells.

FFR is one of the interference management techniques which require minimal or no coordination among the adjacent cells. The basic mechanism of FFR is to partition the macrocell area into spatial regions and each sub-region is assigned different frequency sub-bands for users. In [1, 2], the user satisfaction metric introduced by the authors is evaluated for users' mobility and the performance is compared with other reuse schemes. The authors propose a dynamic mechanism that selects the optimal FFR scheme based on the user satisfaction metric. The proposed mechanism is evaluated through several simulation scenarios that incorporate users' mobility and its selected FFR scheme is compared with other frequency reuse schemes in order to highlight its performance. The research is further enhanced in [3] where cell edge reuse factor is set to 1.5 and results are generated to determine the optimal inner radius and frequency allocation. In [4, 5], a performance study is carried out, where FFR partitions each cell into two regions; inner region and outer region, and allocates different frequency band to each region based on reuse factor. In [6], a frequency reuse technique is proposed which aims at maximizing throughput via combinations of inner cell radius and frequency allocation to the macrocell. In this work, the authors study the interference mitigation techniques in femtocell/macrocell networks and proposed a FFR mechanism that leads to increased overall system performance. Work in [7] proposes a mechanism that selects the optimal FFR scheme based on the user throughput and user satisfaction. In [8, 9], authors present a FFR optimization scheme based on capacity density (bit/s/m^2), which show better performance compared to conventional Reuse-1 and Reuse-3 schemes. The authors formulate an optimization problem and solve it by simulation. Graph theory and similar optimization techniques are presented in [10], a graph-based framework for dynamic FFR in multi-cell OFDMA networks is proposed in this work. The proposed scheme enhances the conventional FFR by

*Corresponding Author Uttara Sawant, University of North Texas, Denton, Texas 76207, +1-940-595-1489 & uttarasawant@my.unt.edu

https://dx.doi.org/10.25046/aj030162

enabling adaptive spectral sharing per cell load conditions. Such adaptation has significant benefits in a practical environment, where traffic load in different cells may be asymmetric and time-varying. Work in [11, 12] provides analysis of the inter cell interference coordination problem in multi-cell OFDMA systems. In order to reduce the interference without losing much frequency resources in each cell, cell users are partitioned into two classes, interior and exterior users. Results determine the optimal configuration of the interior and exterior regions' dimensions as well as the optimal frequency reuse factor.

Authors in [13] use two-stage heuristic approach to find optimal frequency partitioning. The authors intend to propose the FFR scheme by introducing the concept of normalized Spectral Efficiency (nSE). With the optimal Frequency Reuse Factor (FRF), the FFR scheme can maximize the system SE with throughput improvement of cell-edge users. To solve the problem, they divide it into two sub-problems. However, this scheme is asymptotically optimal with the assumption of the uniform distributions of signal and interference in sectors. The simulation results demonstrate the gain of the system SE is about 3% by proposed FFR scheme. The goal in work in [14] is to improve the cell edge throughput as well as the average cell throughput, compared to a network with frequency reuse factor 1. The cell capacity under those reuse schemes is estimated and compared. To attain both high spectral efficiency and good coverage within sectors/beams, a scheme based on coordinated scheduling between sectors of the same site, and the employment of frequency reuse factor above 1 only in outer parts of the sector, is proposed and evaluated. The resulting sector throughput increases with the number of active users. When terminals have one antenna and channels are Rayleigh fading, it results in a sector payload capacity between 1.2 (one user) and 2.1 bits/s/Hz/sector (for 30 users) in an interference-limited environment [15]. In [16], the authors review some of the recent advances in ICI research and discuss the assumptions, advantages, and limitations of the proposed mechanisms. They propose a scheduling algorithm to schedule users based on channel quality and quality of service (QoS) metrics.

Authors in [17] investigate an OFDMA radio resource control (RRC) scheme, where RRC control is exercised at both RNC and base stations. In [18], the authors present an analytical solution to carry out performance study of various frequency reuse schemes in an OFDMA based cellular network. Results of performance study show that results obtained through analytical method are in conformity with those obtained through simulations. In [19], the authors analyze the FFR scheme and propose a dynamic FFR mechanism that selects the optimal frequency allocation based on the cell total throughput and user satisfaction.

Extensive work is done to optimize network performance of FFR mechanism [20]. FFR and exploitation of the channel state information at the transmitter (CSIT) are effective approaches to improving the spectrum efficiency of the outer coverage region. When channels vary within a physical transmission frame, the above improvement is substantially suppressed. To remedy this, work in [21] develops new FFR patterns for OFDMA systems with frequency or time division duplexing (FDD/TDD) in time-varying channels. Simulation results show significant performance gains of the proposed schemes over the existing ones. A semi-centralized joint cell muting and user scheduling scheme for interference coordination in a multi-cell network is proposed under two different temporal fairness criteria. The authors in [22] propose a general pattern set construction algorithm in this paper. Numerical results are provided to validate the effectiveness of the proposed scheme for both criteria. The impact of choice of the cell muting pattern set is also studied through numerical examples for various cellular topologies. Densely deployed cellular wireless networks, which employ small cell technology, are being widely implemented. Authors in [23, 24] aim to maximize the system's throughput through the employment of FFR schemes. The authors derive the optimal configuration of the FFR scheme and evaluate the systems performance behavior under absolute and proportional fairness requirements. Table 1 shows comparison between previous published work and paper contributions.

Table 1: Comparison

Parameter	Previous Work	Current Paper
Throughput variance	High	Low
Femtocells adaptation	Minimal	High

This paper evaluates the performance of the sectorized FFR mechanism with five metrics; four existing and one proposed, which determine the values for inner region radius and frequency allocation for optimal results. For a specific mobility model, an adaptation process is applied to the FFR mechanism optimized for each of the five metrics and its performance is evaluated. Results are also generated and evaluated for non-adaptation process for the same network. Results prove that the proposed metric shows lower variance in user throughput compared to the other metrics from previous published work. The optimized FFR mechanism shows reacting to the adaptation process with user mobility even with the addition of femtocells.

2 FFR Mechanism

The topology of Figure 1 consists of 16 cells with four non-overlapping resource sets. Each cell of the topology is divided into two regions; inner and outer region. The total available bandwidth of the system is split

into four uneven spectrum, denoted by A (blue), B (green), C (red) and D (yellow). Spectrum A, B, and C have equal bandwidth and are allocated in outer regions with Frequency Reuse 3. On the other hand, spectrum D is allocated in all inner regions with Frequency Reuse 1. The frequency resources in all inner regions are universally used, since the inner region users are less exposed to ICI.

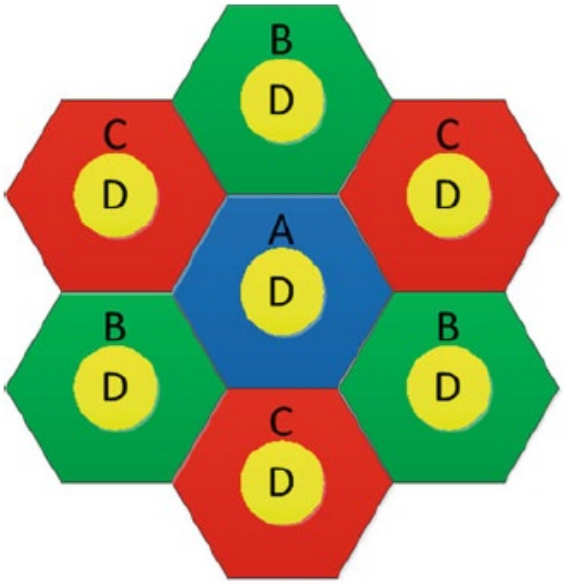


Figure 1: Strict FFR Deployment in LTE Macrocell Deployment [4]

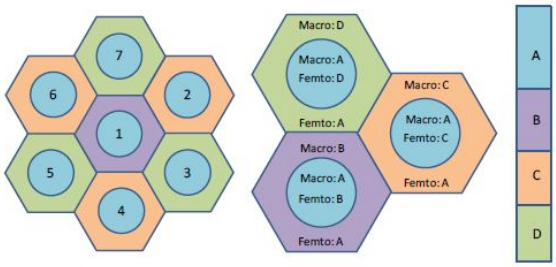


Figure 2: Strict FFR Deployment in LTE Macrocell-Femtocell Deployment [25]

From user's perspective, Integer Frequency Reuse (IFR) can be regarded as a special case of FFR. In IFR, all resource blocks (RBs) allocated to a cell can be used anywhere in the cell without any specification of user's location. For comparison, the FFR scheme that is selected by adaptive mechanism is compared with variations of IFR. The macrocell coverage area is partitioned into center-zone and edge-zone. The entire frequency band is divided into two parts, one part is solely assigned to the center-zone (e.g., sub-band A in Figure 1) and the other part is partitioned into three subbands (e.g., sub-bands B, C, and D) and assigned to the three edge-zones [26]. Figure 2 shows the Strict FFR network layout for integrated macrocell-femtocell deployment. Note that femtocell uses different frequency bands than the overlay macrocell to minimize or eliminate the inter-tier interference in the network.

3 System Model

A set of multicast users are uniformly distributed in the grid of 16 macrocells. In order to find the optimal inner region radius and frequency allocation in the deployment, the mechanism divides each cell into two regions and calculates the total throughput for the following 40 Frequency Allocations (FA), assuming Frequency Reuse 1 and x for inner and the outer regions respectively [26], where x is the frequency reuse factor of 3. Each FA corresponds to paired value of fraction of inner region resource blocks and inner region radius.

- FA1: All 25 resource blocks are allocated in inner region. No resource blocks are allocated to the outer region.
- FA2: 24 resource blocks are allocated in inner region. $1/x$ resource block allocated to the outer region.
  ⋮
- FA39: 1 resource block allocated in inner region. $24/x$ resource blocks allocated to the outer region.
- FA40: No resource blocks are allocated in inner region. $25/x$ resource blocks allocated to the outer region.

For each FA, the mechanism calculates the total throughput, user satisfaction, user fairness, and weighted throughput values based on new metrics. This procedure is repeated for successive inner cell radius (0 to R, where R is the cell radius). The mechanism selects the optimal FFR scheme that maximizes the cell total throughput. This procedure is repeated periodically in order to take into account users' mobility. Therefore, the per-user throughput, the cell total throughput, User Satisfaction (US), and other metrics are calculated in periodic time intervals (the exact time is beyond the scope of this manuscript) and at each time interval, the FFR scheme that maximizes the above parameters is selected. This periodic process is called adaptation [28]. The system model described above can be used for supported LTE bandwidths ranging from 1.4 MHz to 20 MHz.

The signal-to-interference-plus-noise (SINR) for downlink transmission to macro user x on a subcarrier n can be expressed as,

$$SINR_{x,n} = \frac{P_{M,n} G_{x,M,n}}{N_0 \Delta f + \sum_{M'} P_{M',n} G_{x,M',n}} \quad (1)$$

where, $P_{M,n}$ and $P_{M',n}$ is transmit power of serving macrocell M and neighboring macrocell M' on subcarrier n, respectively. $G_{i,M,n}$ is channel gain between macro user i and serving macrocell M on subcarrier n and $G_{i,M',n}$ corresponds to channel gain from neighboring macrocell M'. Finally, N_0 is white noise power spectral density and Δf is subcarrier spacing.

The channel gain G, given by the following equation is dominantly affected by path loss, which is assumed to be modeled based on urban path-loss PL is defined as,

$$G = 10^{-PL/10} \tag{2}$$

Additionally, for the throughput calculation, the capacity of a user i on a specific subcarrier n can be estimated via the $SINR$ from the following equation,

$$C_{i,n} = W \cdot \log_2(1 + \alpha \cdot SINR_{i,n}) \tag{3}$$

where, W denotes the available bandwidth for each subcarrier divided by the number of users that share the specific subcarrier and α is a constant for a target bit error rate (BER) defined by $\alpha = -1.5/ln(5 \cdot BER)$. Here, BER is set to 10^{-6}.

The expression of the total throughput of the serving macrocell M is given as follows,

$$T_i = \sum_n \beta_{i,n} C_{i,n} \tag{4}$$

where, $\beta_{i,n}$ represents the subcarrier assigned to user i. When $\beta_{i,n}$=1, the subcarrier is assigned to user i. Otherwise, $\beta_{i,n}$=0.

4 Mobility Model

This paper evaluates network performance in a scenario with mobile users in the highlighted cell of the topology presented in Figure 3. During the experiment that lasts for 217 seconds, 24 users of the examined cell are moving randomly inside the cell with speed 3 km/h, according to the Pedestrian A channel model [27]. It is assured that all of them remain into the cell's area, ensuring that their total number will remain constant. This corresponds to low mobility scenario with zero to minimal handover.

5 Performance Metrics

The paper evaluates the network performance for adaptive and non-adaptive FFR mechanisms with user mobility. For comparison, the adaptive FFR scheme that is selected by proposed mechanism is compared with three different cases. The first case, where the optimal inner radius and frequency allocation remain constant through the adaptation process, is referred to FFR non-adaptive. The second case, where the cell bandwidth equals the whole network bandwidth, is called IFR with frequency reuse 1 (IFR1). In the third case, the inner region radius is zero and each cell uses one third of the networks bandwidth. This case is called IFR with frequency reuse 3 (IFR3). The difference with the IFR1 case, lies in the fact that only co-channel base station are considered in interference calculation and as a consequence, the interference base station density is divided by 3.

5.1 Existing Metrics

This paper uses a metric US defined by authors in [28]. It is calculated as the sum of the users' throughput divided by the product of the maximum user's throughput and the number of users (X). US ranges between 0 and 1. When US approaches 1, all the users in the corresponding cell experience similar throughput. However, when US approaches 0, there are huge differences in throughput values across the users in the cell. This metric will be utilized in scenarios where fairness to the users is significant such as cell throughput and Jain fairness index.

$$US = \frac{\sum_{x=1}^{X} T_x}{max\_user\_throughput \cdot X} \tag{5}$$

With metric WT defined by the authors in [28], the aim is to not only generate low variance of the per-user throughput values but also obtain higher values of the cell total throughput.

$$WT = JI \cdot T \tag{6}$$

where T is the cell mean throughput.

To obtain a metric of fairness for performance evaluation, the Jain fairness metric introduced in [29] is used. Assuming the allocated throughput for user i is x_i, Jain fairness index for the cell is defined as,

$$JI = \frac{(\sum_{i=1}^{X} x_i)^2}{X \cdot \sum_{i=1}^{X} x_i^2} \tag{7}$$

This metric is interesting for the evaluation of the proposed method due to its properties. It is scale-independent, applicable for different number of users and it is bounded between [0, 1], where 0 means "total unfairness" and 1 means "total fairness" in terms of throughput division among the users.

5.2 Proposed Metric

This section introduces a new metric, weighted throughput based on user satisfaction WT_{US}, to add weights to the cell throughput corresponding to specific inner radius and inner bandwidth such that the resultant throughput is higher than the corresponding throughput optimized at user satisfaction alone and all the users in the corresponding cell experience similar throughput.

$$WT_{US} = US \cdot T \tag{8}$$

From (6), (8), it is clear that the metrics are cell-based.

6 Performance Evaluation

6.1 Mathematical Model

The product model is applied as mathematical model in the new metric definition and development. Previous metric, weighted throughput, adds the benefits

of the individual component metrics, throughput and fairness index. Similarly, the proposed metric defined by the product model of throughput and user satisfaction, is expected to perform better than the individual component metrics.

6.2 Simulation Parameters

Table 2 are parameters set for simulation.

Table 2: Simulation Parameters

Parameter	Value
System Bandwidth	5 MHz
Subcarriers	300
Subcarrier Bandwidth	180 KHz
Cell Radius	250 m
Inter eNodeB distance	500 m
Noise Power Spectral Density	-174 dBm/Hz
Subcarrier spacing	15 KHz
Channel Model	Typical Urban
Carrier Frequency	2000 MHz
Number of macrocells	16
Macrocell Transmit Power	46 dBm
Path Loss	Cost 231 Hata Model
Users' speed	3 km/h PedA

A sample uniform deployment considered in simulation is shown in Figure 3. Macrocells are located at cell centers. Active users are randomly distributed and are shown with white dots.

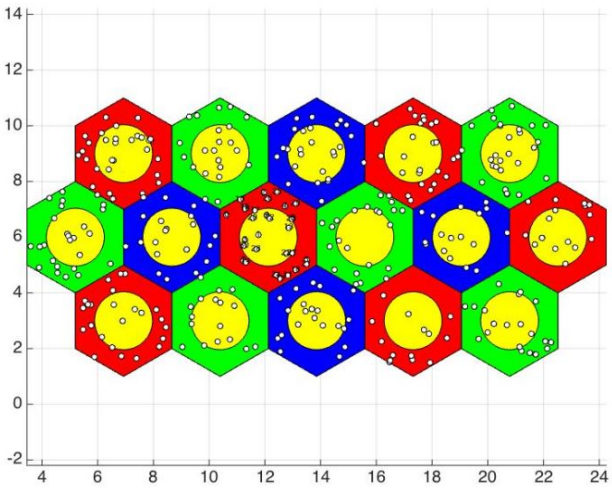


Figure 3: Strict FFR Uniform Deployment for Simulation

6.3 Simulation Guidelines

Note that the simulation is run assuming a stable downlink data traffic, since the simulation framework is analyzing data for downlink traffic. A network snapshot where user association to macrocell remains stable for the duration of the simulation run is considered as the users are associated to the base stations with maximum SINR. A frequent handover will occur in high mobility environment and that is not included in the scope of this paper. The simulation software for our research is based on [30].

6.4 Performance Analysis

6.4.1 Comparison of New Metric to Other Metrics

The new metric is introduced due to the following reasons, better performance in terms of average user throughput and effective user fairness in terms of variance in user throughput. Figure 4 proves the user throughput optimized with new metric shows better performance and low variance.

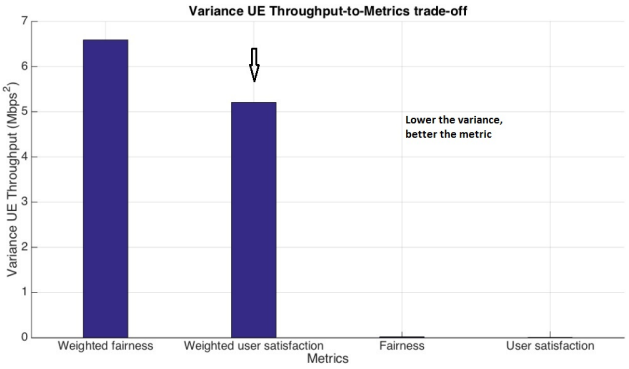


Figure 4: Comparison of Variance in User Throughput

6.4.2 Static Users - Metric Performance

Simulation is carried out by applying the metrics on users with static positions and performance is evaluated with respect to inner region subcarriers and inner region radius. Figures 5 - 8 show the metric performance with respect to inner region subcarriers. Figures 9 - 12 show the metric performance with respect to inner region radius.

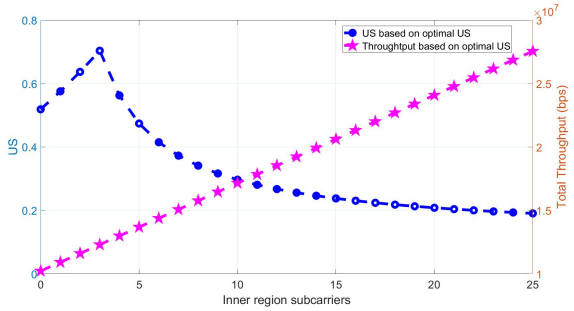


Figure 5: User Satisfaction Metric Performance for Static Users to Inner Region Subcarriers

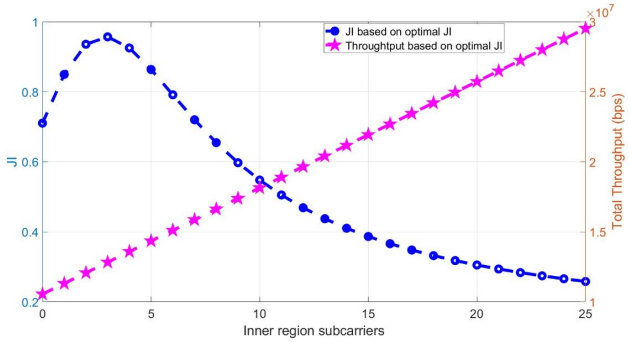


Figure 6: Fairness Index Metric Performance for Static Users to Inner Region Subcarriers

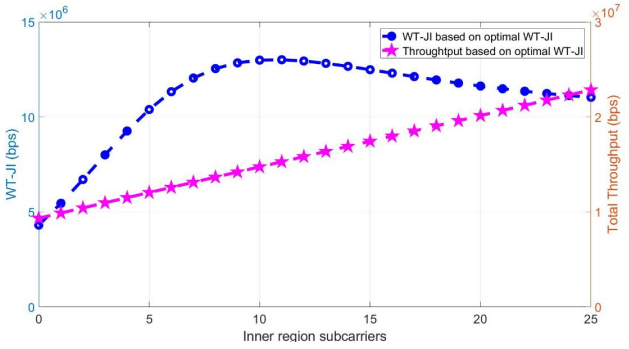


Figure 7: Weighted Fairness Metric Performance for Static Users to Inner Region Subcarriers

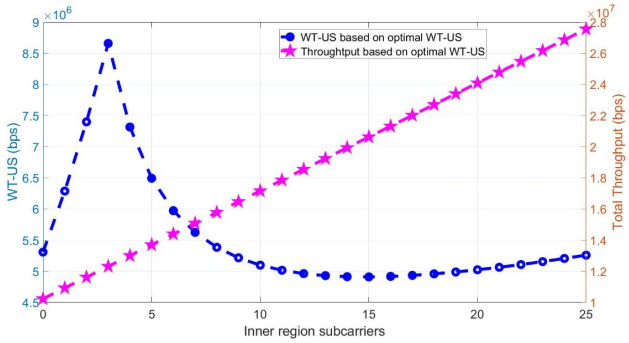


Figure 8: Weighted User Satisfaction Metric Performance for Static Users to Inner Region Subcarriers

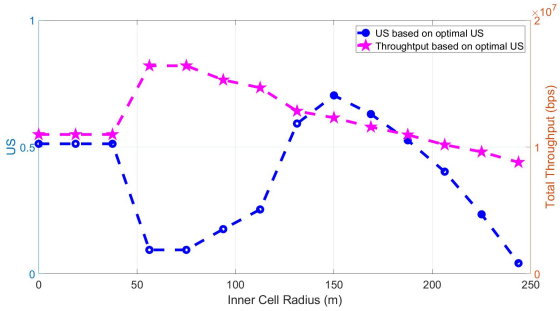


Figure 9: User Satisfaction Metric Performance for Static Users to Inner Region Radius

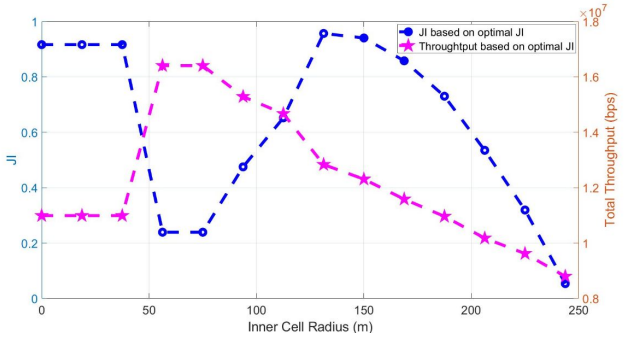


Figure 10: Fairness Index Metric Performance for Static Users to Inner Region Radius

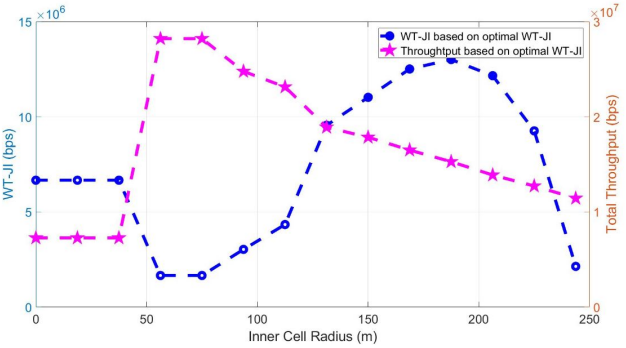


Figure 11: Weighted Fairness Metric Performance for Static Users to Inner Region Radius

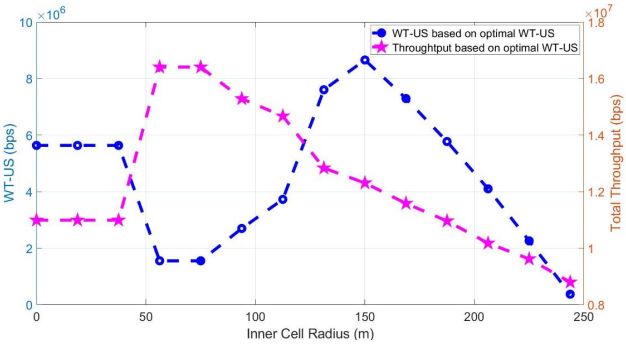


Figure 12: Weighted User Satisfaction Metric Performance for Static Users to Inner Region Radius

6.4.3 Adaptive versus Non-adaptive mechanisms - Metric Performance

As user mobility is introduced and new positions are determined, the simulation calculates metrics and optimal inner radius and subcarrier allocation for adaptive and non-adaptive variations for each FFR mechanism. For non-adaptive FFR mechanism, optimal inner radius and subcarrier allocation remain constant even after the user position changes and performance metrics are calculated. For adaptive FFR mechanism, new values of optimal inner radius and subcarrier allocation are determined based on the new user position and optimal FFR performance. Therefore, the adaptive FFR mechanism reacts to user mobility better than non-adaptive FFR mechanism. Figures 13 - 17 show

FFR performance for all metrics with adaptive and non-adaptive mechanisms applied. In all the adaptive versus non-adaptive mechanisms comparison results, the adaptive mechanism applied on the metrics perform better than non-adaptive mechanism. Particularly, the proposed metric performance Figure 17 shows higher throughput than other metrics and better performance versus non-adaptive and IFR variations. For both weighted user satisfaction (proposed metric) in Figure 17 and weighted fairness (existing metric) in Figure 16, the adaptation process is visible between 50 to 200 seconds. However, despite common benefit of better reaction to the adaptation process, weighted user satisfaction performs better with higher throughput compared to weighted fairness. There is a co-relation between adaptation trend and inner region radius as shown in Section 6.4.4.

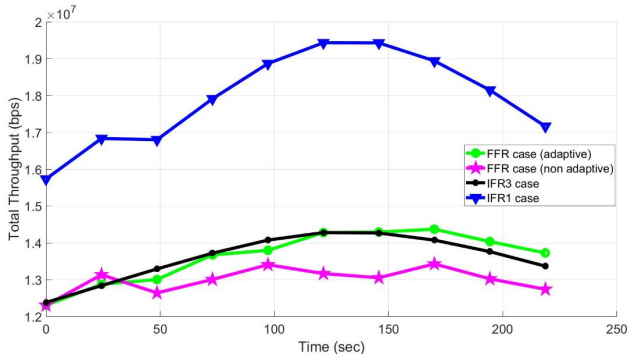


Figure 13: Comparison of adaptive versus non-adaptive mechanisms for Total Throughput

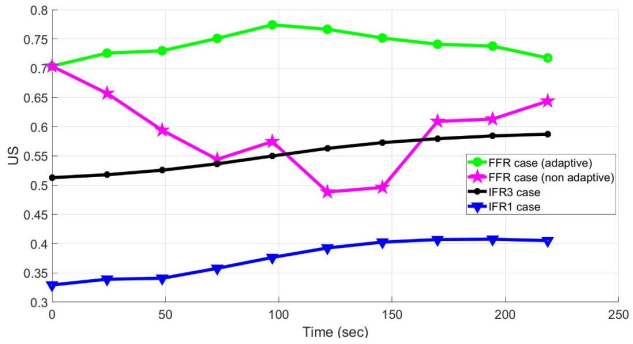


Figure 14: Comparison of adaptive versus non-adaptive mechanisms for User Satisfaction

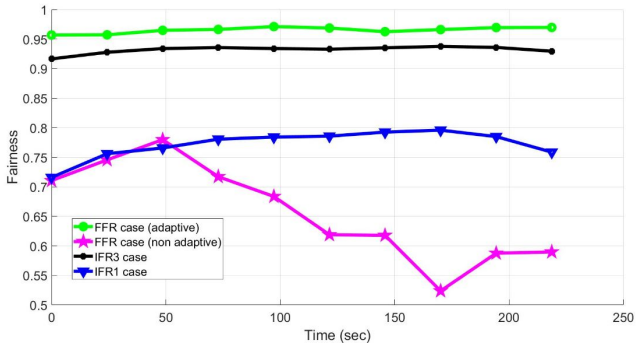


Figure 15: Comparison of adaptive versus non-adaptive mechanisms for Fairness Index

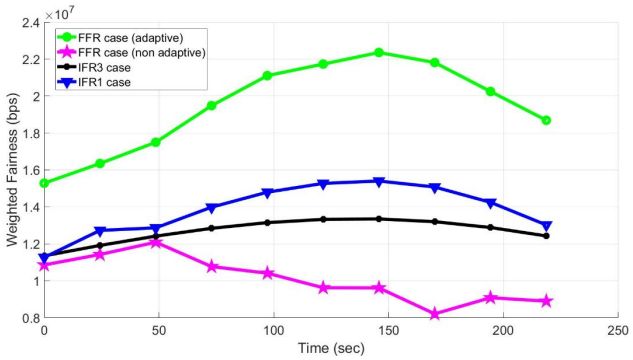


Figure 16: Comparison of adaptive versus non-adaptive mechanisms for Weighted Fairness

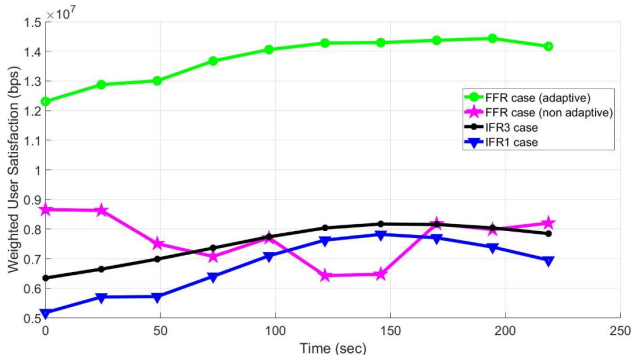


Figure 17: Comparison of adaptive versus non-adaptive mechanisms for Weighted User Satisfaction

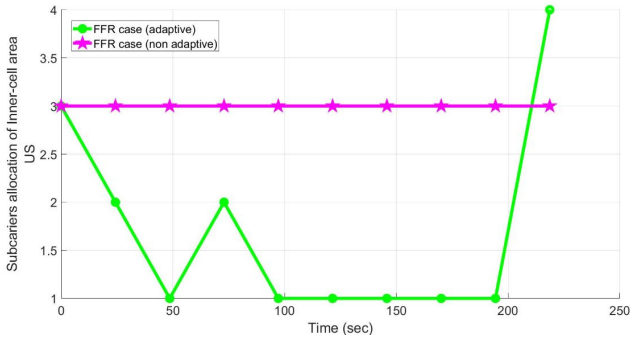


Figure 18: Comparison of Subcarrier Allocation for User Satisfaction

6.4.4 Adaptive versus Non-adaptive mechanisms - Subcarrier Allocation and Inner-Cell Radius

The throughput trend optimized for each metric can be explained with inner region subcarrier and inner region radius results. Figures 22 - 25 show effect of adaptation process on subcarrier allocation for all metrics. Figures 22 - 25 show effect of adaptation process on inner region radius for all metrics. Both subcarrier allocation and inner region radius show active response to the adaptation process applied to the network during the simulation time. For weighted user satisfaction, Figure 25 and weighted fairness Figure 24, the subcarrier allocation reacts positively to the adaptation process between 50 and 200 seconds. High throughput trend during this time interval can be explained due to allocation of all 25 subcarriers. Similarly, the range of

inner region radius changes with the adaptation process during the simulation time compared to its static trend in non-adaptive mechanism as shown in Figures 25, 24.

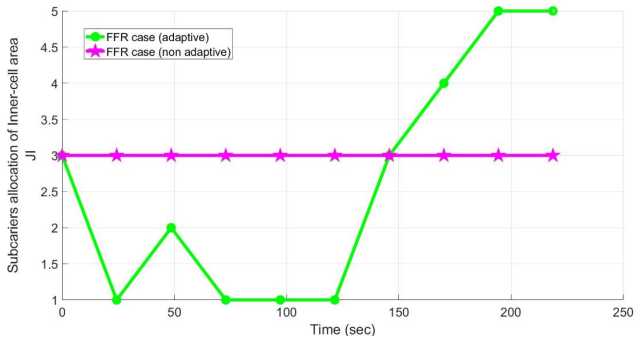


Figure 19: Comparison of Subcarrier Allocation for Fairness Index

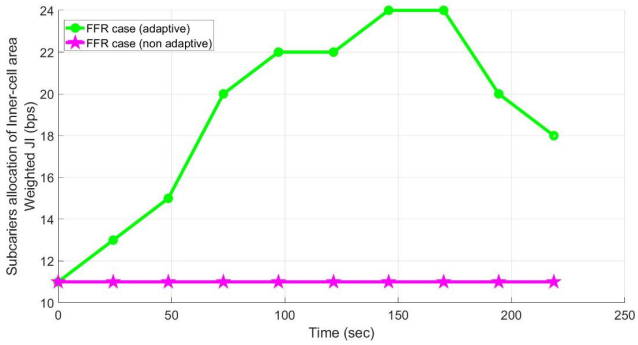


Figure 20: Comparison of Subcarrier Allocation for Weighted Fairness

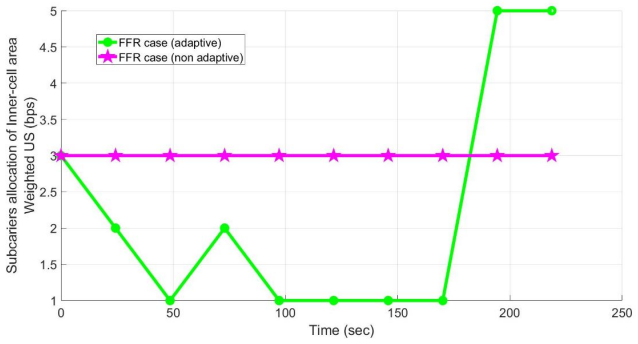


Figure 21: Comparison of Subcarrier Allocation for Weighted User Satisfaction

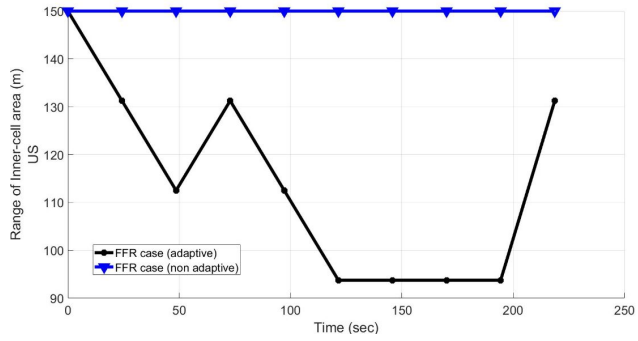


Figure 22: Comparison of Range of Inner-cell for User Satisfaction

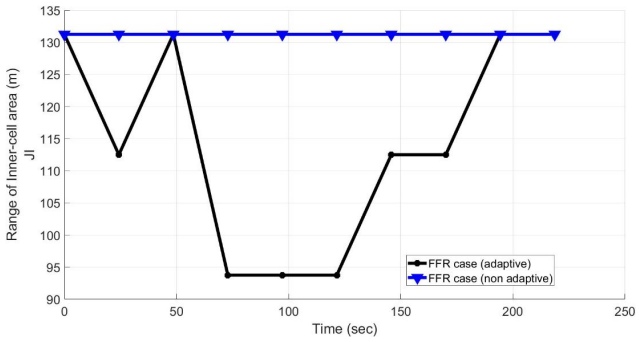


Figure 23: Comparison of Range of Inner-cell for Fairness Index

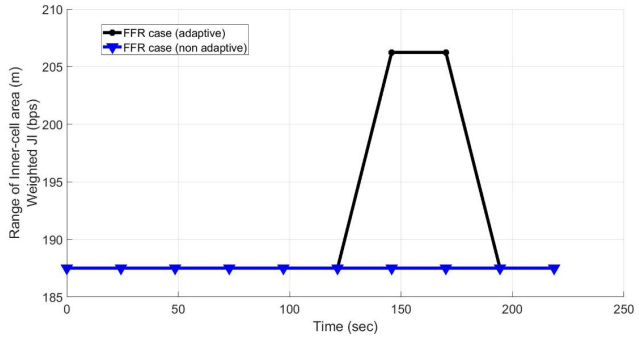


Figure 24: Comparison of Range of Inner-cell for Weighted Fairness

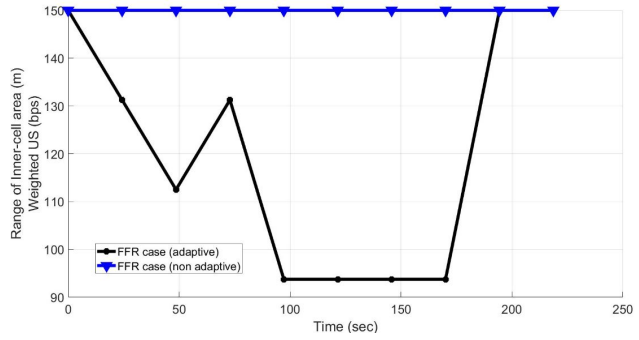


Figure 25: Comparison of Range of Inner-cell for Weighted User Satisfaction

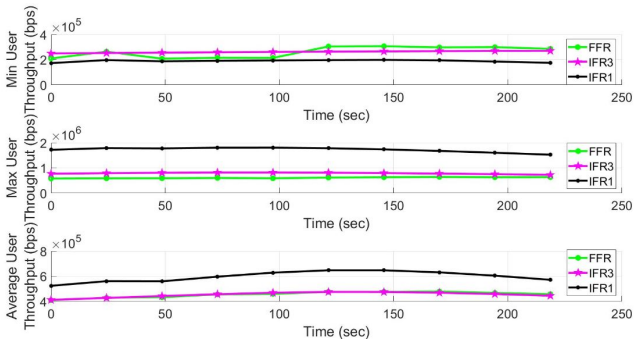


Figure 26: Comparison of Minimum, Maximum, Average for User Satisfaction

6.4.5 Adaptive versus Non-adaptive mechanisms - Minimum, Maximum, and Average Values

FFR mechanism simulation compares their minimum, maximum, and average results over time. Figures 26 - 29 show the metric performance. Maximum and average throughput shows lowest performance for FFR mechanism optimized for weighted user satisfaction metric. This is obvious result since the goal of the new metric is to reduce variance in user throughput and maintain relatively high overall throughput.

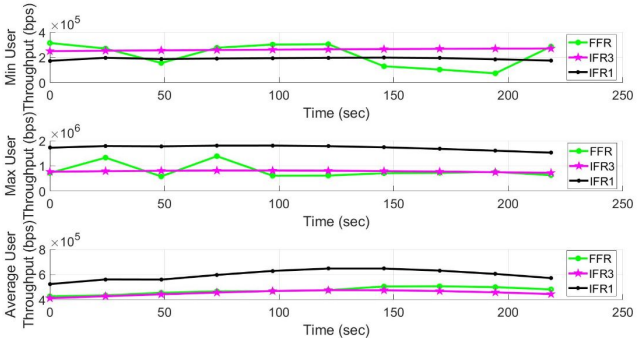


Figure 27: Comparison of Minimum, Maximum, Average for Fairness Index

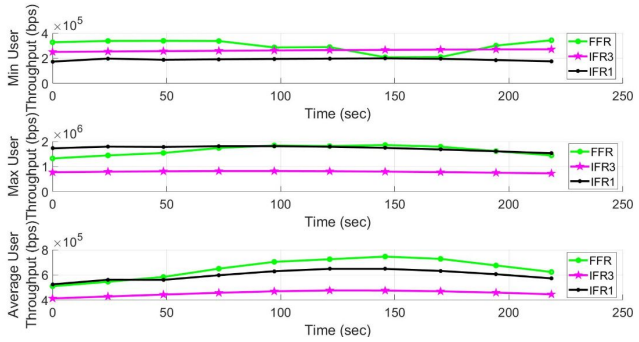


Figure 28: Comparison of Minimum, Maximum, Average for Weighted Fairness

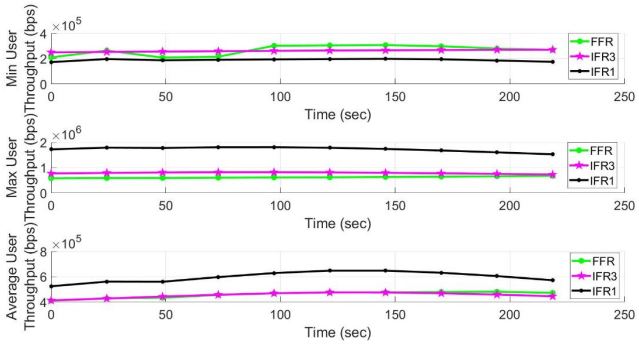


Figure 29: Comparison of Minimum, Maximum, Average for Weighted User Satisfaction

6.4.6 Adaptive versus Non-adaptive mechanisms - Femtocells

LTE femtocells have been developed to increase capacity and raise the throughput of cell-edge users. In this simulation, the LTE network is extended with randomly placed femtocells in each macrocell, assuming an uniform femtocell co-channel deployment. Despite the high throughput benefits, femtocell deployment experience relatively high level of interference from neighboring macrocells and femtocells. FFR frequency deployment is set as indicated in Figure 2. Simulation results in Figure 30 and 31 indicate higher throughput due to addition of femtocell radio resources. Adaptive trend is in line with the previous results without femtocells, where the network throughput adapts to the moving user locations. Network throughput optimized to weighted user satisfaction metric shows higher maximum and average throughput compared to IFR1 and IFR3 due to the presence of additional frequency resources.

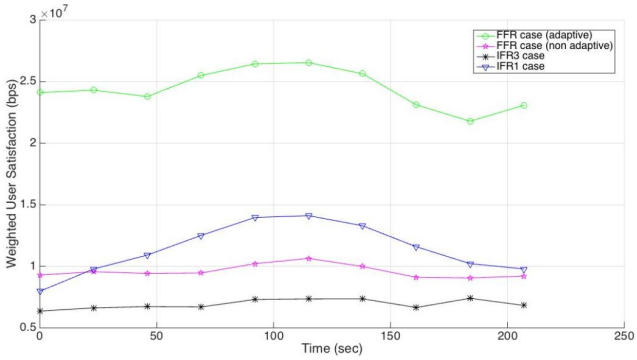


Figure 30: Comparison of adaptive versus non-adaptive mechanisms for Weighted User Satisfaction with Femtocells

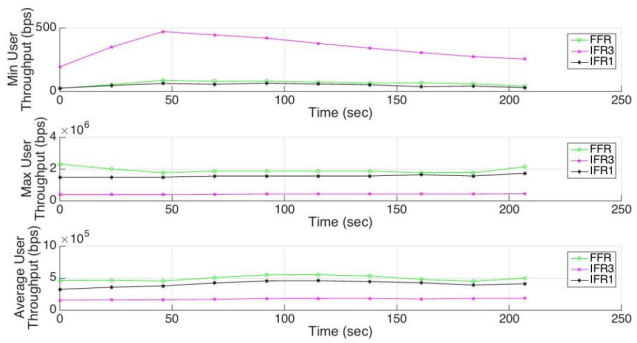


Figure 31: Comparison of Minimum, Maximum, Average for Weighted User Satisfaction with Femtocells

7 Conclusion

In conclusion, the paper evaluated the adaptation process in LTE FFR mechanism. The selected FFR mechanism based on the optimal inner region radius and frequency allocation performed better than other interference co-ordination mechanisms. The FFR mechanism optimized using weighted throughput user satisfaction made a positive trade-off between the existing metrics, as it maintained relatively high total throughput and reduced the variance of per-user throughput. The proposed metric reacted better to the adaptation process in mobility scenarios and generated higher throughput compared to the other metrics. With extreme densification in future mobility, the effect of femtocells on

the network throughput optimized for new metric on moving users was studied. A good research extension would be to review impact of femtocell densification and uniform grid placement on the adaptation process with the proposed metric.

Conflict of Interest The authors declare no conflict of interest.

References

[1] U. Sawant, R. Akl, "Evaluation of Adaptive and Non Adaptive LTE Fractional Frequency Reuse Mechanisms" in The 26th Annual Wireless and Optical Communications Conference, New Jersey, USA, 2017. https://doi.org/10.1109/WOCC.2017.7928998

[2] D. Bilios, C. Bouras, V. Kokkinos, G. Tseliou, A. Papazois, "Selecting the Optimal Fractional Frequency Reuse Scheme in Long Term Evolution Networks" in IEEE Wireless Pers. Communications, 2013. https://doi.org/10.1007/s11277-012-0965-z

[3] M. Yadav, M. Palle, A. Hani, "Performance Analysis of Fractional Frequency Reuse Factor for Interference Suppression in Long Term Evolution" in International Journal of Conceptions on Electronics and Communication Engineering, 2015. https://doi.org/10.1109/PGSRET.2015.7312249

[4] C. Bouras, D. Bilios, V. Kokkinos, A. Papazois, G. Tseliou, "A performance study of Fractional Frequency Reuse in OFDMA networks" in IEEE Wireless Telecommunications Symposium, 2012. https://doi.org/10.1109/WMNC.2012.6416137

[5] C. Bouras, G. Kavourgias, V. Kokkinos, A. Papazois, "Interference management in LTE femtocell systems using an adaptive frequency reuse scheme" in IEEE Wireless and Mobile Networking Conference (WMNC), 2012. https://doi.org/10.1109/PGSRET.2015.7312249

[6] C. Bouras, G. Kavourgias, V. Kokkinos, A. Papazois, "Interference management in LTE femtocell systems using an adaptive frequency reuse scheme" in IEEE Wireless Telecommunications Symposium, 2012. https://doi.org/10.1109/WTS.2012.6266120

[7] D. Bilios, C. Bouras, V. Kokkinos, A. Papazois, G. Tseliou, "Optimization of fractional frequency reuse in Long Term Evolution networks" in IEEE Wireless Communications and Networking Conference (WCNC), 2012. https://doi.org/10.1109/WCNC.2012.6214087

[8] M. Taranetz, J. Ikuno, M. Rupp, "Capacity Density Optimization by Fractional Frequency Partitioning" in IEEE, 2011. https://doi.org/10.1109/ACSSC.2011.6190246

[9] J. Ikuno, M. Taranetz, M. Rupp, "A Fairness-based Performance Evaluation of Fractional Frequency Reuse in LTE" in 2013 17th International ITG Workshop on Smart Antennas (WSA), 2013.

[10] Y. Chang, Z. Tao, J. Zhang, C. Kuo, "A graph approach to dynamic fractional frequency reuse (FFR) in multi-cell OFDMA networks" in In proceedings of IEEE international conference on communications (ICC 2009), 2009. https://doi.org/10.1109/ICC.2009.5198612

[11] M. Assad, "Optimal fractional frequency reuse (FFR) in multicellular OFDMA system" in In proceedings of IEEE 68th Vehicular technology conference (VTC 2008Fall), 2008. https://doi.org/10.1109/VETECF.2008.381

[12] N. Hassan, M. Assad, "Optimal fractional frequency reuse (FFR) and resource allocation in multiuser OFDMA system" in In Proceedings of International Conference on Information and Communication Technologies, (ICICT 2009), 2009. https://doi.org/10.1109/ICICT.2009.5267207

[13] L. Fang, X. Zhang, "Optimal fractional frequency reuse in OFDMA based wireless networks" in Proceedings of 4th international conference on wireless communications, networking and mobile computing, (WiCOM08), 2008. https://doi.org/10.1109/WiCom.2008.166

[14] Y. Xiang, J. Luo, C. Hartmann, "Inter-cell interference mitigation through flexible resource reuse in OFDMA based communication networks" in Proceedings of European Wireless, 2007.

[15] M. Sternad, T. Ottosson, A. Ahlen, A. Svensson, "Attaining both coverage and high spectral efficiency with adaptive OFDM downlinks" in Proceedings of IEEE 58th Vehicular Technology Conference (VTC 2003-Fall), 2003. https://doi.org/10.1109/VETECF.2003.1285981

[16] G. Fodor, C. Koutsimanis, A. Rcz, N. Reider, A. Simonsson, W. Mller, "Intercell interference coordination in OFDMA networks and in the 3GPP long term evolution system" in Journal of Communications, 2009. https://doi.org/10.4304/jcm.4.7.445-453

[17] G. Li, H. Liu, "Downlink radio resource allocation for multi-cell OFDMA system" in IEEE Transactions on Wireless Communications, 2006. https://doi.org/10.1109/TWC.2006.256968

[18] P. Godlewski, M. Maqbool, M. Coupechoux, J. Kelif, "Analytical evaluation of various frequency reuse schemes in cellular OFDMA networks" in Proceedings of 3rd international conference on performance evaluation methodologies and tools (Valuetools 2008), 2008. https://doi.org/10.1016/j.peva.2009.08.001

[19] J. Lim, R. Badlishah, M. Jusoh, "LTE-fractional frequency reuse (FFR) optimization with femtocell network" in 2nd International Conference on Electronic Design (ICED), 2014. https://doi.org/10.1109/ICED.2014.7015863

[20] H. Elfadil, M. Ali, M. Abas, "Fractional frequency reuse in LTE networks" in 2015 2nd World Symposium on Web Applications and Networking (WSWAN), 2015. https://doi.org/10.1109/WSWAN.2015.7210297

[21] P. Yen, Q. Zhan, H. Minn, "New Fractional Frequency Reuse Patterns for Multi-Cell Systems in Time-Varying Channels" in International Journal of Conceptions on Electronics and Communication Engineering, 2015. https://doi.org/10.1109/LWC.2015.2404787

[22] S. Shahsavari, N. Akar, B. Hossein Khalaj, "Joint Cell Muting and User Scheduling in Multi-Cell Networks with Temporal Fairness" in CoRR, 2016.

[23] H. Chang, I. Rubin, "Optimal Downlink and Uplink Fractional Frequency Reuse in Cellular Wireless Networks" in IEEE Transactions on Vehicular Technology, 2015. https://doi.org/10.1109/TVT.2015.2425356

[24] I. Mahmoud, O. Elgzzar, S. Hashima, "An accurate model of worst case signal to interference ratio for frequency reuse cellular systems" in 2016 11th International Conference on Computer Engineering and Systems (ICCES), 2016. https://doi.org/10.1109/ICCES.2016.7822037

[25] N. Saquib, E. Hossain, D. I. Kim, "Fractional frequency reuse for interference management in LTE-Advanced HetNets", in IEEE Wireless Communications, 2013. https://doi.org/10.1109/MWC.2013.6507402

[26] P. Lee, T. Lee, J. Jeong, J. Shin, "Interference Management in LTE Femtocell Systems Using Fractional Frequency Reuse" in The 12th International Conference on Advanced Communication Technology (ICACT), 2010. https://doi.org/10.1109/PGSRET.2015.7312249

[27] Technical Specification Group RAN, "Technical Specification Group RAN, *E-UTRA; LTE RF system scenarios*" in 3GPP Tech. Rep. TS 36.942, 2008. https://doi.org/10.1109/PGSRET.2015.7312249

[28] C. Bouras, V. Kokkinos, A. Papazois, G. Tseliou, "Fractional Frequency Reuse in Integrated Femtocell/Macrocell Environments" in IEEE WWIC, 2013. https://doi.org/10.1007/978-3-642-38401-1_18

[29] R. Jain, D. Chiu, W. Hawe, "A Quantitative Measure of Fairness And Discrimination for Resource Allocation in Shared Computer System" in DEC Technical Report 301, 1984. https://doi.org/10.1109/PGSRET.2015.7312249

[30] Research Unit 6, "FFR Scheme Selection Mechanism", 2013. http://ru6.cti.gr/ru6/FFR_selection_mechanism.zip

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 1, 521-536 (2018)
www.astesj.com
Special issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Two-Stage Performance Engineering of Container-based Virtualization

Zheng Li[1,4], Maria Kihl[*,2], Yiqun Chen[3], He Zhang[1]

[1]*Software Institute, Nanjing University, 210008, China*

[2]*Department of Electrical and Information Technology, Lund University, 223 63, Sweden*

[3]*Centre for Spatial Data Infrastructures & Land Administration, University of Melbourne, 3010, Australia*

[4]*Department of Computer Science, University of Concepción, 4070386, Chile*

ARTICLEINFO

Article history:
Received: 14 November, 2017
Accepted: 05 February, 2018
Online: 28 February, 2018

Keywords:
Cloud Computing
Container
Hypervisor
MapReduce
Performance Engineering
Virtualization

ABSTRACT

Cloud computing has become a compelling paradigm built on compute and storage virtualization technologies. The current virtualization solution in the Cloud widely relies on hypervisor-based technologies. Given the recent booming of the container ecosystem, the container-based virtualization starts receiving more attention for being a promising alternative. Although the container technologies are generally considered to be lightweight, no virtualization solution is ideally resource-free, and the corresponding performance overheads will lead to negative impacts on the quality of Cloud services. To facilitate understanding container technologies from the performance engineering's perspective, we conducted two-stage performance investigations into Docker containers as a concrete example. At the first stage, we used a physical machine with "just-enough" resource as a baseline to investigate the performance overhead of a standalone Docker container against a standalone virtual machine (VM). With findings contrary to the related work, our evaluation results show that the virtualization's performance overhead could vary not only on a feature-by-feature basis but also on a job-to-job basis. Moreover, the hypervisor-based technology does not come with higher performance overhead in every case. For example, Docker containers particularly exhibit lower QoS in terms of storage transaction speed. At the ongoing second stage, we employed a physical machine with "fair-enough" resource to implement a container-based MapReduce application and try to optimize its performance. In fact, this machine failed in affording VM-based MapReduce clusters in the same scale. The performance tuning results show that the effects of different optimization strategies could largely be related to the data characteristics. For example, LZO compression can bring the most significant performance improvement when dealing with text data in our case.

1 Introduction

The container technologies have widely been accepted for building next-generation Cloud systems. This paper investigates the performance overhead of container-based virtualization and the performance optimization of a container-based MapReduce application, which is an extension of work originally presented in the 31st IEEE International Conference on Advanced Information Networking and Application (AINA 2017) [1].

The Cloud has been considered to be able to provide computing capacity as the next utility in our modern daily life. In particular, it is the virtualization technologies that enable Cloud computing to be a new paradigm of utility, by playing various vital roles in supporting Cloud services, ranging from resource iso-

[*]Corresponding Author; Address: Ole Römers Väg 3, Lund 223 63, Sweden; Tel: + 46 46 222 9010; Email: maria.kihl@eit.lth.se

https://dx.doi.org/10.25046/aj030163

lation to resource provisioning. The existing virtualization technologies can roughly be distinguished between the hypervisor-based and the container-based solutions. Considering their own resource consumption, both virtualization solutions inevitably introduce performance overheads to running Cloud services, and the performance overheads could then lead to negative impacts to the corresponding quality of service (QoS). Therefore, it would be crucial for both Cloud providers (e.g., for improving infrastructural efficiency) and consumers (e.g., for selecting services wisely) to understand to what extend a candidate virtualization solution incurs influence on the Cloud's QoS.

Recall that hypervisor-driven virtual machines (VMs) require guest operating systems (OS), while containers can share a host OS. Suppose physical machines, VMs and containers are three candidate resource types for a particular Cloud service, a natural hypothesis could be:

> *The physical machine-based service has the best quality among the three resource types, while the container-based service performs better than the hypervisor-based VM service.*

Unfortunately, to the best of our knowledge, there is little quantitative evidence to help test this hypothesis in an "apple-to-apple" manner, except for the similar qualitative discussions. Furthermore, the performance overhead of hypervisor-based and container-based virtualization technologies can even vary in practice depending on different service circumstance (e.g., uncertain workload densities and resource competitions). Therefore, we decided to conduct a twofold investigation into containers from the performance's perspective. Firstly, we used a physical machine with "just-enough" resource as a baseline to quantitatively investigate and compare the performance overheads between the container-based and hypervisor-based virtualizations. In particular, since Docker is currently the most popular container solution [2] and VMWare is one of the leaders in the hypervisor market [3], we chose Docker and VMWare Workstation 12 Pro to represent the two virtualization solutions respectively. Secondly, we implemented a container-based MapReduce cluster on a physical machine with "fair-enough" resource to investigate the performance optimization of our MapReduce application at least in this use case.

According to the clarifications in [4, 5], our qualitative investigations can be regulated by the discipline of experimental computer science (ECS). By employing ECS's recently available Domain Knowledge-driven Methodology (DoKnowMe) [6], we experimentally explored the performance overheads of different virtualization solutions on a feature-by-feature basis, i.e. the communication-, computation-, memory- and storage-related QoS aspects. As for the investigation into performance optimization, we were concerned with the task timeout, out-of-band heartbeat, buffer setting, stream merging, data compression and the cluster size of our container-based MapReduce application.

The experimental results and analyses of performance overhead investigation generally advocate the aforementioned hypothesis. However, the hypothesis is not true in all the cases. For example, we do not see computation performance difference between the three resource types for solving a combinatorially hard chess problem; and the container exhibits even higher storage performance overhead than the VM when reading/writing data byte by byte. Moreover, we find that the remarkable performance loss incurred by both virtualization solutions usually appears in the performance variability.

The performance optimization investigation reveals that various optimization strategies might take different effects due to different data characteristics of a container-based MapReduce application. For example, dealing with text data can significantly benefit from enabling data compression, whereas buffer settings have little effect for dealing with relatively small amount of data.

Overall, our work makes fourfold contributions to the container ecosystem, as specified below.

(1) Our experimental results and analyses can help both researchers and practitioners to better understand the fundamental performance of the present container-based and hypervisor-based virtualization technologies. In fact, the performance evaluation practices in ECS can roughly be distinguished between two stages: the first stage is to reveal the primary performance of specific (system) features, while the second stage is generally based on the first-stage evaluation to investigate real-world application cases. Thus, this work can be viewed as a foundation for more sophisticated evaluation studies in the future.

(2) Our method of calculating performance overhead can easily be applied or adapted to different evaluation scenarios by others. The literature shows that the "performance overhead" has normally been used in the context of qualitative discussions. By quantifying such an indicator, our study essentially provides a concrete lens into the case of performance comparisons.

(3) As a second-stage evaluation work, our case study on the performance optimization of a MapReduce application both demonstrates a practical use case and supplies an easy-to-replicate scenario for engineering performance of container-based applications. In other words, this work essentially proposed a characteristic-consistent data set (i.e. Amazon's spot price history that is open to the public) for future performance engineering studies.

(4) The whole evaluation logic and details reported in this paper can be viewed as a reusable template of evaluating Docker containers. Since the Docker project is still quickly growing [7], the evaluation results could be gradually out of date. Given this template, future evaluations

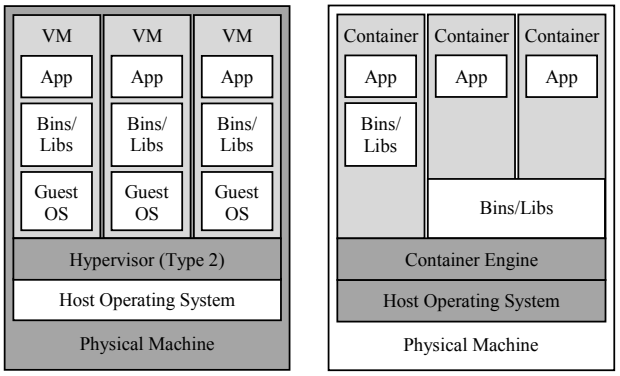
can be conveniently repeated or replicated even by different evaluators at different times and locations. More importantly, by emphasizing the backend logic and evaluation activities, the template-driven evaluation implementations (instead of results only) would be more traceable and comparable.

The remainder of this paper is organized as follows. Section 2 briefly summarizes the background knowledge of container-based and the hypervisor-based virtualization technologies. Section 3 introduces the fundamental performance evaluation of a single container. The detailed performance overhead investigation is divided into two reporting parts, namely pre-experimental activities and experimental results & analyses, and they are correspondingly described into Section 3.2 and 3.3 respectively. Section 4 explains our case study on the performance optimization of a container-based MapReduce application. Section 5 highlights the existing work related to container's performance evaluation. Conclusions and some future work are discussed in Section 6.

2 Hypervisor-based vs. Container-based Virtualization

When it comes to the Cloud virtualization, the de facto solution is to employ the hypervisor-based technologies, and the most representative Cloud service type is offering VMs [8]. In this virtualization solution, the hypervisor manages physical computing resources and makes isolated slices of hardware available for creating VMs [7]. We can further distinguish between two types of hypervisors, namely the bare-metal hypervisor that is installed directly onto the computing hardware, and the hosted hypervisor that requires a host OS. To make a better contrast between the hypervisor-related and container-related concepts, we particularly emphasize the second hypervisor type, as shown in Figure 1a. Since the hypervisor-based virtualization provides access to physical hardware only, each VM needs a complete implementation of a guest OS including the binaries and libraries necessary for applications [9]. As a result, the guest OS will inevitably incur resource competition against the applications running on the VM service, and essentially downgrade the QoS from the application's perspective. Moreover, the performance overhead of the hypervisor would also be passed on to the corresponding Cloud services and lead to negative impacts on the QoS.

To relieve the performance overhead of hypervisor-based virtualization, researchers and practitioners recently started promoting an alternative and lightweight solution, namely container-based virtualization. In fact, the foundation of the container technology can be traced back to the Unix `chroot` command in 1979 [9], while this technology is eventually evolved into virtualization mechanisms like Linux VServer, OpenVZ and Linux Containers (LXC) along with the booming of Linux [10]. Unlike the hardware-level solution of hypervisors, containers realize virtualization at the OS level and utilize isolated slices of the host OS to shield their contained applications [9]. In essence, a container is composed of one or more lightweight images, and each image is a prebaked and replaceable file system that includes necessary binaries, libraries or middlewares for running the application. In the case of multiple images, the read-only supporting file systems are stacked on top of each other to cater for the writable top-layer file system [2]. With this mechanism, as shown in Figure 1b, containers enable applications to share the same OS and even binaries/libraries when appropriate. As such, compared to VMs, containers would be more resource efficient by excluding the execution of hypervisor and guest OS, and more time efficient by avoiding booting (and shutting down) a whole OS [11, 7]. Nevertheless, it has been identified that the cascading layers of container images come with inherent complexity and performance penalty [12]. In other words, the container-based virtualization technology could also negatively impact the corresponding QoS due to its performance overhead.



(a) Hypervisor-based virtual service. (b) Container-based virtual service.

Figure 1: Different architectures of hypervisor-based and container-based virtual services.

3 Fundamental Performance Evaluation of a Single Container

3.1 Performance Evaluation Methodology

Since the comparison between the container's and the VM's performance overheads is essentially based on their performance evaluation, we define our work as a performance evaluation study that belongs to the field of experimental computer science [4, 5]. Considering that "evaluation methodology underpins all innovation in experimental computer science" [13], we employ the methodology DoKnowMe [6] to guide evaluation implementations in this study. DoKnowMe is an abstract evaluation methodology on the analogy of "class" in object-oriented programming. By integrating domain-specific knowledge artefacts, DoKnowMe can be customized into specific methodologies (by analogy

of "object") to facilitate evaluating different concrete computing systems. The skeleton of DoKnowMe is composed of ten generic evaluation steps, as listed below.

(1) Requirement recognition;
(2) Service feature identification;
(3) Metrics and benchmarks listing;
(4) Metrics and benchmarks selection;
(5) Experimental factors listing;
(6) Experimental factors selection;
(7) Experiment design;
(8) Experiment implementation;
(9) Experimental analysis;
(10) Conclusion and documentation.

Each evaluation step further comprises a set of activities together with the corresponding evaluation strategies. The elaboration on these evaluation steps is out of the scope of this paper. To better structure our report, we divide the evaluation implementation into pre-experimental activities and experimental results & analyses.

3.2 Pre-Experimental Activities

3.2.1 Requirement Recognition

Following DoKnowMe, the whole evaluation implementation is essentially driven by the recognized requirements. In general, the requirement recognition is to define a set of specific requirement questions both to facilitate understanding the real-world problem and to help achieve clear statements of the corresponding evaluation purpose. In this case, the basic requirement is to give a fundamental quantitative comparison between the hypervisor-based and container-based virtualization solutions. As mentioned previously, we concretize these two virtualization solutions into VMWare Workstation VMs and Docker containers respectively, in order to facilitate our evaluation implementation (i.e., using a physical machine as a baseline to investigate the performance overhead of a Docker container against a VM). Thus, such a requirement can further be specified into two questions:

RQ1: How much performance overhead does a standalone Docker container introduce over its base physical machine?

RQ2: How much performance overhead does a standalone VM introduce over its base physical machine?

Considering that virtualization technologies could lead to variation in performance of Cloud services [14], we are also concerned with the container's and VM's potential variability overhead besides their average performance overhead:

RQ3: How much performance variability overhead does a standalone Docker container introduce over its base physical machine during a particular period of time?

RQ4: How much performance variability overhead does a standalone VM introduce over its base physical machine during a particular period of time?

3.2.2 Service Feature Identification

Recall that we treat Docker containers as an alternative type of Cloud service to VMs. By using the taxonomy of Cloud services evaluation [15], we directly list the service feature candidates, as shown in Figure 2. Note that a service feature is defined as a combination of a physical property and its capacity, and we individually examine the four physical properties in this study:

- Communication
- Computation
- Memory
- Storage

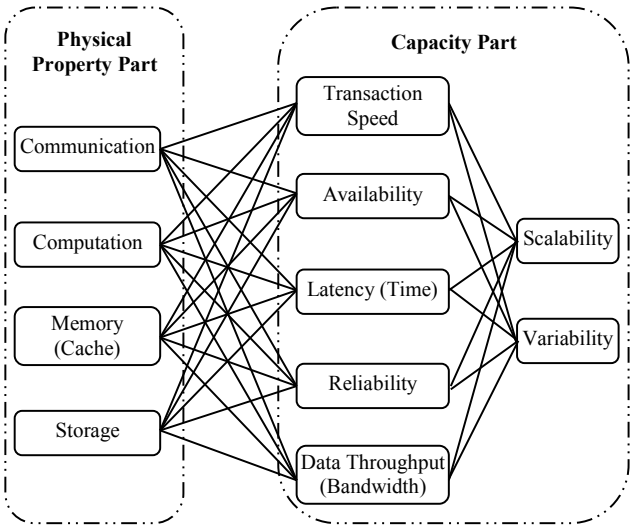


Figure 2: Candidate service features for evaluating Cloud service performance (cf. [15]).

3.2.3 Metrics/Benchmarks Listing and Selection

The selection of evaluation metrics usually depends on the availability of benchmarks. According to our previous experience of Cloud services evaluation, we choose relatively lightweight and popular benchmarks to try to minimize the benchmarking overhead, as listed in Table 1. For example, Iperf has been identified to be able to deliver more precise results by consuming less system resources. In fact, except for STREAM that is the de facto memory evaluation benchmark included in the HPC Challenge Benchmark (HPCC) suite, the other benchmarks are all Ubuntu's built-in utilities.

Physical Property	Capacity Metric	Benchmark	Version
Communication	Data Throughput	Iperf	2.0.5
Computation	(Latency) Score	HardInfo	0.5.1
Memory	Data Throughput	STREAM	5.10
Storage	Transaction Speed	Bonnie++	1.97.1
Storage	Data Throughput	Bonnie++	1.97.1

Table 1: Metrics and benchmarks for this evaluation study.

In particular, although Bonnie++ only measures the amount of data processed per second, the disk I/O transactions are on a byte-by-byte basis when accessing small size of data. Therefore, we consider to measure storage transaction speed when operating byte-size data and measure storage data throughput when operating block-size data. As for the property computation, considering the diversity in CPU jobs (e.g., integer and floating-point calculations), we employ HardInfo that includes six micro-benchmarks to generate performance scores, as briefly explained in Table 2. HardInfo is a tool package that can summarize the information about the host machine's hardware and operating system, as well as benchmarking the CPU. In this study, we employ HardInfo for CPU benchmarking only.

Benchmark	Brief Explanation
CPU Blowfish	Encrypting blocks of random data using the Blowfish algorithm.
CPU CryptoHash*	Checking the ability of the computer to find the hash of a specific test file.
CPU Fibonacci	Calculating the 42*nd* Fibonacci number.
CPU N-Queens	Solving the combinatorially hard chess problem of placing N queens on an $N \times N$ chessboard such that no queen can attack any other.
FPU FFT	Computing a fast Fourier transform.
FPU Raytracing	Generating an image by tracing the path of light through pixels in an image plane and simulating the effects of its encounters with virtual objects.

*The higher the better. (The lower the better for the others.)

Table 2: Micro-benchmarks included in HardInfo.

When it comes to the performance overhead, we use the business domain's *Overhead Ratio*[1] as an analogy to its measurement. In detail, we treat the performance loss compared to a baseline as the expense, while imagining the baseline performance to be the overall income, as defined in Equation (1).

$$O_p = \frac{|P_m - P_b|}{P_b} \times 100\% \quad (1)$$

where O_p refers to the performance overhead; P_m denotes the benchmarking result as a measurement of a service feature; P_b indicates the baseline performance of the service feature; and then $|P_m - P_b|$ represents the corresponding performance loss. Note that the physical machine's performance is used as the baseline in our study. Moreover, considering possible observational errors, we allow a margin of error for the confidence level as high as 99% with regarding to the benchmarking results. In other words, we will ignore the difference between the measured performance and its baseline if the calculated performance overhead is less than 1% (i.e. if $O_p < 1\%$, then $P_m = P_b$).

[1] http://www.investopedia.com/terms/o/overhead-ratio.asp

3.2.4 Experimental Factor Listing and Selection

The identification of experimental factors plays a prerequisite role in the following experimental design. More importantly, specifying the relevant factors would be necessary for improving the repeatability of experimental implementations. By referring to the experimental factor framework of Cloud services evaluation [16], we choose the resource- and workload-related factors as follows.

The resource-related factors:

- *Resource Type:* Given the evaluation requirement, we have essentially considered three types of resources to support the imaginary Cloud service, namely physical machine, container and VM.
- *Communication Scope:* We test the communication between our local machine and an Amazon EC2 t2.micro instance. The local machine is located in our broadband lab at Lund University, and the EC2 instance is from Amazon's available zone ap-southeast-1a within the region Asia Pacific (Singapore).
- *Communication Ethernet Index:* Our local side uses a Gigabit connection to the Internet, while the EC2 instance at remote side has the "Low to Moderate" networking performance defined by Amazon.
- *CPU Index:* The physical machine's CPU model is Intel Core™2 Duo Processor T7500. The processor has two cores with the 64-bit architecture, and its base frequency is 2.2 GHz. We allocate both CPU cores to the standalone VM upon the physical machine.
- *Memory Size:* The physical machine is equipped with a 3GB DDR2 SDRAM. When running the VMWare Workstation Pro without launching any VM, "`watch -n 5 free -m`" shows a memory usage of 817MB while leaving 2183MB free in the physical machine. Therefore, we set the memory size to 2GB for the VM to avoid (at least to minimize) the possible memory swapping.
- *Storage Size:* There are 120GB of hard disk in the physical machine. Considering the space usage by the host operating system, we allocate 100GB to the VM.
- *Operating System:* Since Docker requires a 64-bit installation and Linux kernels older than

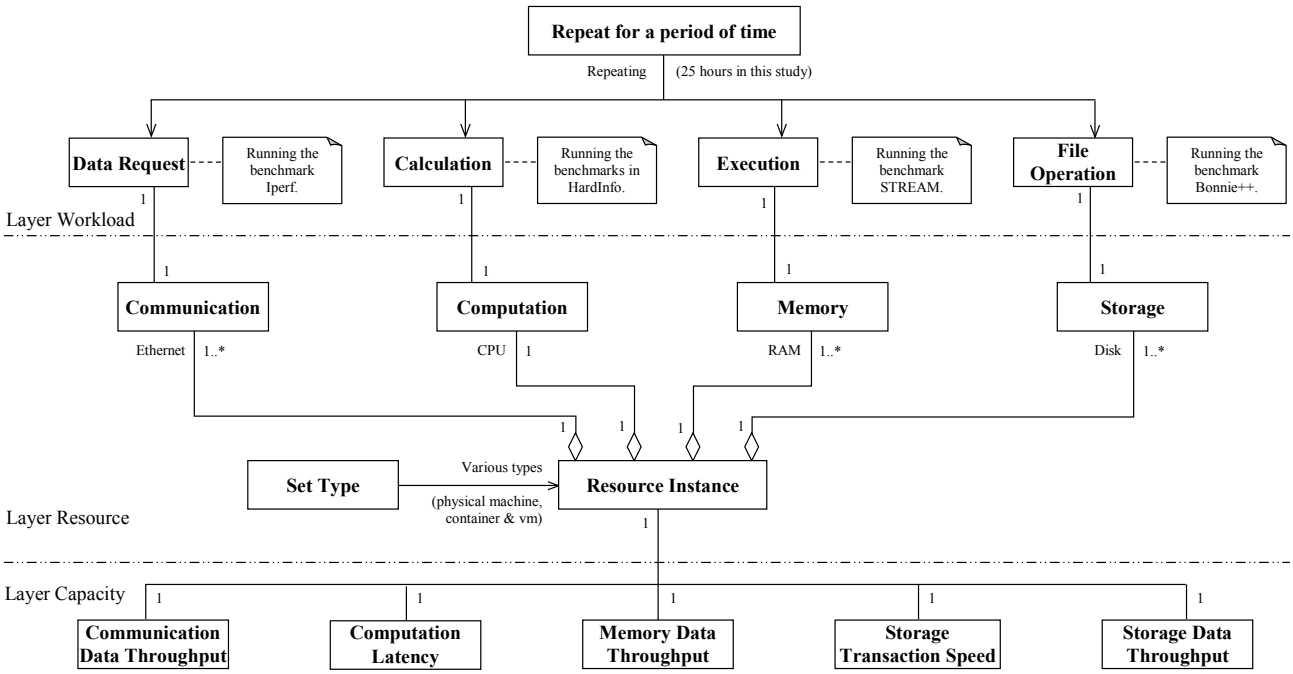


Figure 3: Experimental blueprint for evaluating three types of resources in this study.

3.10 do not support all the features for running Docker containers, we choose the latest 64-bit Ubuntu 15.10 as the operating system for both the physical machine and the VM. In addition, according to the discussions about base images in the Docker community [17, 18], we intentionally set an OS base image (by specifying `FROM ubuntu:15.10` in the Dockerfile) for all the Docker containers in our experiments. Note that a container's OS base image is only a file system representation, while not acting as a guest OS.

The workload-related factors:

- *Duration:* For each evaluation experiment, we decided to take a whole-day observation plus one-hour warming up (i.e. 25 hours).
- *Workload Size:* The experimental workloads are predefined by the selected benchmarks. For example, the micro-benchmark CPU Fibonacci generates workload by calculating the 42*nd* Fibonacci number (cf. Table 2). In particular, the benchmark Bonnie++ distinguishes between reading/writing byte-size and block-size data.

3.2.5 Experimental Design

It is clear that the identified factors are all with single value except for the *Resource Type*. Therefore, a straightforward design is to run the individual benchmarks on each of the three types of resources independently for a whole day plus one hour.

Furthermore, following the conceptual model of IaaS performance evaluation [19], we record the experimental design into a blueprint both to facilitate our experimental implementations and to help other evaluators replicate/repeat our study. In particular, the experimental elements are divided into three layers (namely Layer Workload, Layer Resource, and Layer Capacity), as shown in Figure 3. To avoid duplication, we do not elaborate the detailed elements in this blueprint.

3.3 Experimental Results and Analyses

3.3.1 Communication Evaluation Result and Analysis

Docker creates a virtual bridge `docker0` on the host machine to enable both the host-container and the container-container communications. In particular, it is the Network Address Translation (NAT) that forwards containers' traffic to external networks. For the purpose of "apple-to-apple" comparison, we also configure VM's network type as NAT that uses the `VMnet8` virtual switch created by VMware Workstation.

Using NAT, both Docker containers and VMs can establish outgoing connections by default, while they require port binding/forwarding to accept incoming connections. To reduce the possibility of configurational noise, we only test the outgoing communication performance, by setting the remote EC2 instance to Iperf server and using the local machine, container and VM all as Iperf clients.

The benchmarking results of repeating `iperf -c XXX.XXX.XXX.XXX -t 15` (with a one-minute interval between every two consecutive trials) are listed in Table 3. The `XXX.XXX.XXX.XXX` denotes the external IP address of the EC2 instance used in our experiments. Note that, unlike the other performance features, the communication data throughput delivers periodical and significant fluctuations, which might be a result from the network resource competition at both our local side and the EC2 side during working hours. There-

fore, we particularly focus on the longest period of relatively stable data out of the whole-day observation, and thus the results here are for rough reference only.

Resource Type	Average	Standard Deviation
Physical machine	29.066 Mbits/sec	1.282 Mbits/sec
Container	28.484 Mbits/sec	1.978 Mbits/sec
Virtual machine	12.843 Mbits/sec	2.979 Mbits/sec

Table 3: Communication benchmarking results using Iperf.

Given the extra cost of using the NAT network to send and receive packets, there would be unavoidable performance penalties for both the container and the VM. Using Equation (1), we calculate their communication performance overheads, as illustrated in Figure 4.

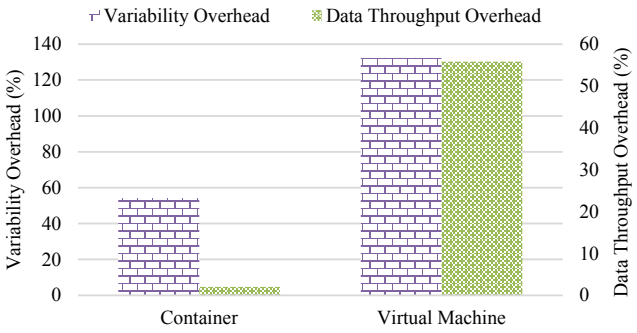


Figure 4: Communication data throughput and its variability overhead of a standalone Docker container vs. VM (using the benchmark Iperf).

A clear trend is that, compared to the VM, the container loses less communication performance, with only 2% data throughput overhead and around 54% variability overhead. However, it is surprising to see a more than 55% data throughput overhead for the VM. Although we have double checked the relevant configuration parameters and redone several rounds of experiments to confirm this phenomenon, we still doubt about the hypervisor-related reason behind such a big performance loss. We particularly highlight this observation to inspire further investigations.

3.3.2 Computation Evaluation Result and Analysis

Recall that HardInfo's six micro benchmarks deliver both "higher=better" and "lower=better" CPU scores (cf. Table 2). To facilitate experimental analysis, we use the two equations below to standardize the "higher=better" and "lower=better" benchmarking results respectively.

$$HB_i = \frac{Benchmarking_i}{\max(Benchmarking_{1,2,\ldots,n})} \quad (2)$$

$$LB_i = \frac{\frac{1}{Benchmarking_i}}{\max\left(\frac{1}{Benchmarking_{1,2,\ldots,n}}\right)} \quad (3)$$

where HB_i further scores the service resource type i by standardizing the "higher=better" benchmarking result $Benchmarking_i$; and similarly, LB_i represents the standardized "lower=better" CPU score of the service resource type i. Note that Equation (3) essentially offers the "lower=better" benchmarking results a "higher=better" representation through reciprocal standardization.

For the purpose of conciseness, here we only specify the standardized experimental results, as shown in Table 4. Exceptionally, the container and VM have slightly higher CPU N-Queens scores than the physical machine. Given the predefined observational margin of error (cf. Section 3.2.3), we are not concerned with this trivial difference, while treating their performance values as equal to each other in this case.

Benchmark	Physical machine	Container	VM
CPU Blowfish	1	0.986	0.942
CPU CryptoHash	1	0.992	0.943
CPU Fibonacci	1	0.999	0.976
CPU N-Queens	0.996	1	0.997
FPU FFT	1	0.966	0.924
FPU Raytracing	1	0.968	0.941

Table 4: Standardized computation benchmarking results using HardInfo.

We can further use a radar plot to help ignore the trivial number differences, and also help intuitively contrast the performance of the three resource types, as demonstrated in Figure 5. For example, the different polygon sizes clearly indicate that the container generally computes faster than the VM, although the performance differences are on a case-by-case basis with respect to different CPU job types.

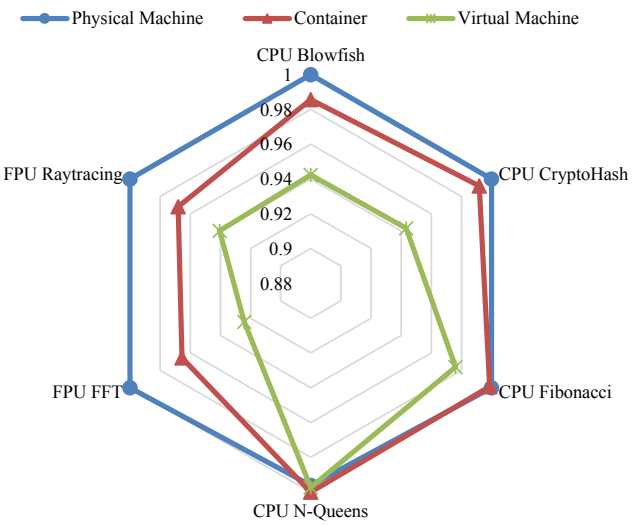


Figure 5: Computation benchmarking results by using HardInfo.

Nevertheless, our experimental results do not display any general trend in variability of those resources' computation scores. As can be seen from the calculated performance overheads (cf. Figure 6), the VM

does not even show worse variability than the physical machine when running CPU CryptoHash, CPU N-Queens and FPU Raytracing. On the contrary, there is an almost 2500% variability overhead for the VM when calculating the 42*nd* Fibonacci number. In particular, the virtualization technologies seem to be sensitive to the Fourier transform jobs (the benchmark FPU FFT), because the computation latency overhead and the variability overhead are relatively high for both the container and the VM.

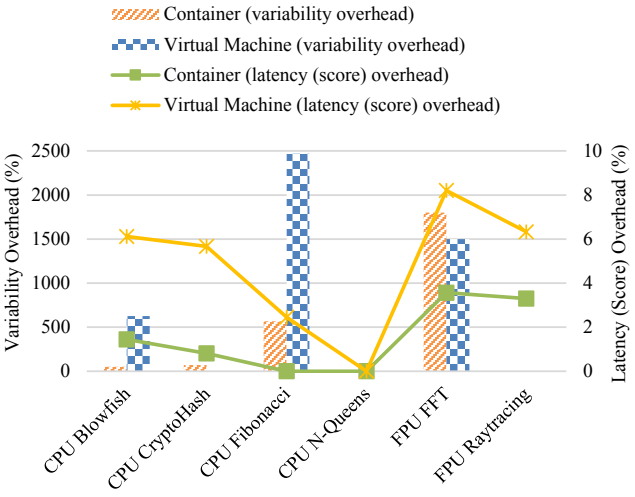


Figure 6: Computation latency (score) and its variability overhead of a standalone Docker container vs. VM (using the tool kit HardInfo).

3.3.3 Memory Evaluation Result and Analysis

STREAM measures sustainable memory data throughput by conducting four typical vector operations, namely Copy, Scale, Add and Triad. The memory benchmarking results are listed in Table 5. We further visualize the results into Figure 7 to facilitate our observation. As the first impression, it seems that the VM has a bit poorer memory data throughput, and there is little difference between the physical machine and the Docker container in the context of running STREAM.

Operation (MB/s)	Physical machine	Container	VM
Copy	2902.685	2914.023	2818.291
(Std. Dev.)	(4.951)	(12.579)	(57.633)
Scale	2916.247	2910.485	2765.737
(Std. Dev.)	(3.783)	(14.488)	(59.193)
Add	3335.634	3332.822	3159.188
(Std. Dev.)	(3.765)	(14.405)	(58.385)
Triad	3341.327	3340.416	3204.811
(Std. Dev.)	(3.976)	(26.361)	(59.004)

Table 5: Memory benchmarking results using STREAM.

By calculating the performance overhead in terms of memory data throughput and its variability, we are able to see the significant difference among these three types of resources, as illustrated in Figure 8. Take the operation Triad as an example, although the container performs as well as the physical machine on average, the variability overhead of the container is more than 500%; similarly, although the VM's Triad data throughput overhead is around 4% only, its variability overhead is almost 1400%. In other words, the memory performance loss incurred by both virtualization techniques is mainly embodied with the increase in the performance variability.

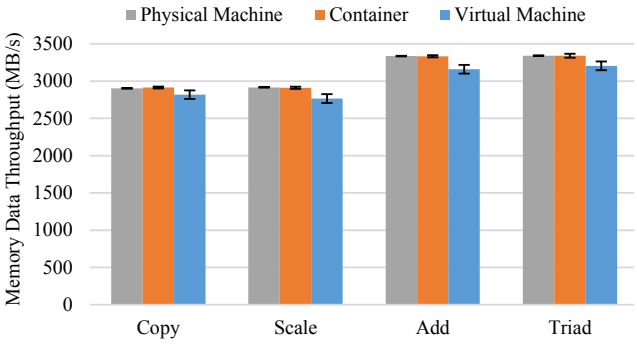


Figure 7: Memory benchmarking results by using STREAM. Error bars indicate the standard deviations of the corresponding memory data throughput.

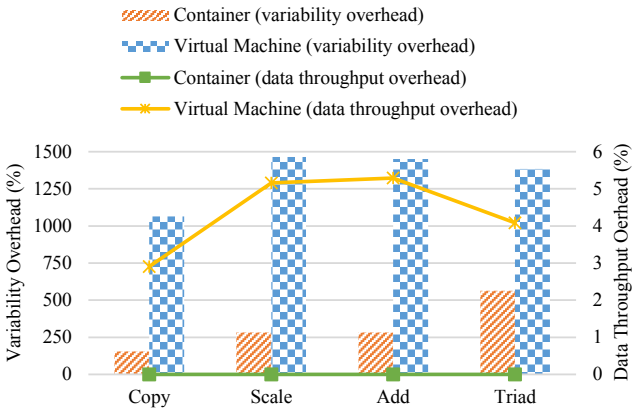


Figure 8: Memory data throughput and its variability overhead of a standalone Docker container vs. VM (using the benchmark STREAM).

In addition, it is also worth notable that the container's average Copy data throughput is even slightly higher than the physical machine (i.e. 2914.023MB/s vs. 2902.685MB/s) in our experiments. Recall that we have considered a 1% margin of error. Since those two values are close to each other within this error margin, here we ignore such an irregular phenomenon as an observational error.

3.3.4 Storage Evaluation Result and Analysis

For the test of disk reading and writing, Bonnie++ creates a dataset twice the size of the involved RAM memory. Since the VM is allocated 2GB of RAM, we also restrict the memory usage to 2GB for Bonnie++

on both the physical machine and the container, by running "`sudo bonnie++ -r 2048 -n 128 -d / -u root`". Correspondingly, the benchmarking trials are conducted with 4GB of random data on the disk. When Bonnie++ is running, it carries out various storage operations ranging from data reading/writing to file creating/deleting. Here we only focus on the performance of reading/writing byte- and block-size data.

To help highlight several different observations, we plot the trajectory of the experimental results along the trial sequence during the whole day, as shown in Figure 9. The first surprising observation is that, all the three resource types have regular patterns of performance jitter in block writing, rewriting and reading. Due to the space limit, we do not report their block rewriting performance in this paper. By exploring the hardware information, we identified the hard disk drive (HDD) model to be ATA Hitachi HTS54161, and its specification describes "It stores 512 bytes per sector and uses four data heads to read the data from two platters, rotating at 5,400 revolutions per minute". As we know, the hard disk surface is divided into a set of concentrically circular tracks. Given the same rotational speed of an HDD, the outer tracks would have higher data throughput than the inner ones. As such, those regular patterns might indicate that the HDD heads sequentially shuttle between outer and inner tracks when consecutively writing/reading block data during the experiments.

The second surprising observation is that, unlike most cases in which the VM has the worst performance, the container seems significantly poor at accessing the byte size of data, although its performance variability is clearly the smallest. We further calculate the storage performance overhead to deliver more specific comparison between the container and the VM, and draw the results into Figure 10. Note that, in the case when the container's/VM's variability is smaller than the physical machine's, we directly set the corresponding variability overhead to zero rather than allowing any performance overhead to be negative. Then, the bars in the chart indicate that the storage variability overheads of both virtualization technologies are nearly negligible except for reading byte-size data on the VM (up to nearly 200%). The lines show that the container brings around 40% to 50% data throughput overhead when performing disk operations on a byte-by-byte basis. On the contrary, there is relatively trivial performance loss in VM's byte data writing. However, the VM has roughly 30% data throughput overhead in other disk I/O scenarios, whereas the container barely incurs overhead when reading/writing large size of data.

Our third observation is that, the storage performance overhead of different virtualization technologies can also be reflected through the total number of the iterative Bonnie++ trials. As pointed by the maximum x-axis scale in Figure 9, the physical machine, the container and the VM can respectively finish 150, 147 and 101 rounds of disk tests during 24 hours. Given this information, we estimate the container's and the VM's storage performance overhead to be 2% (= |147−150|/150) and 32.67% (= |101−150|/150) respectively.

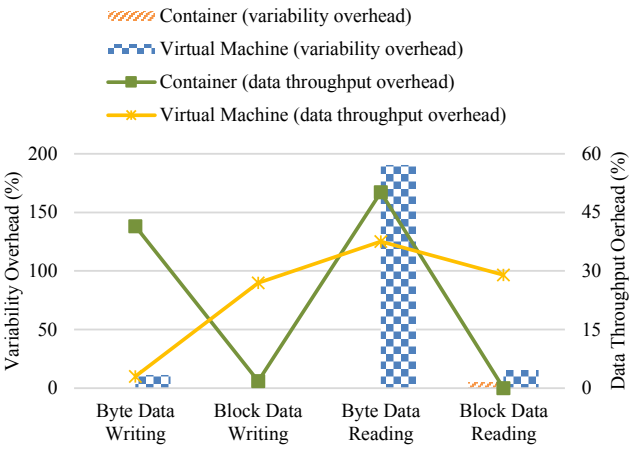
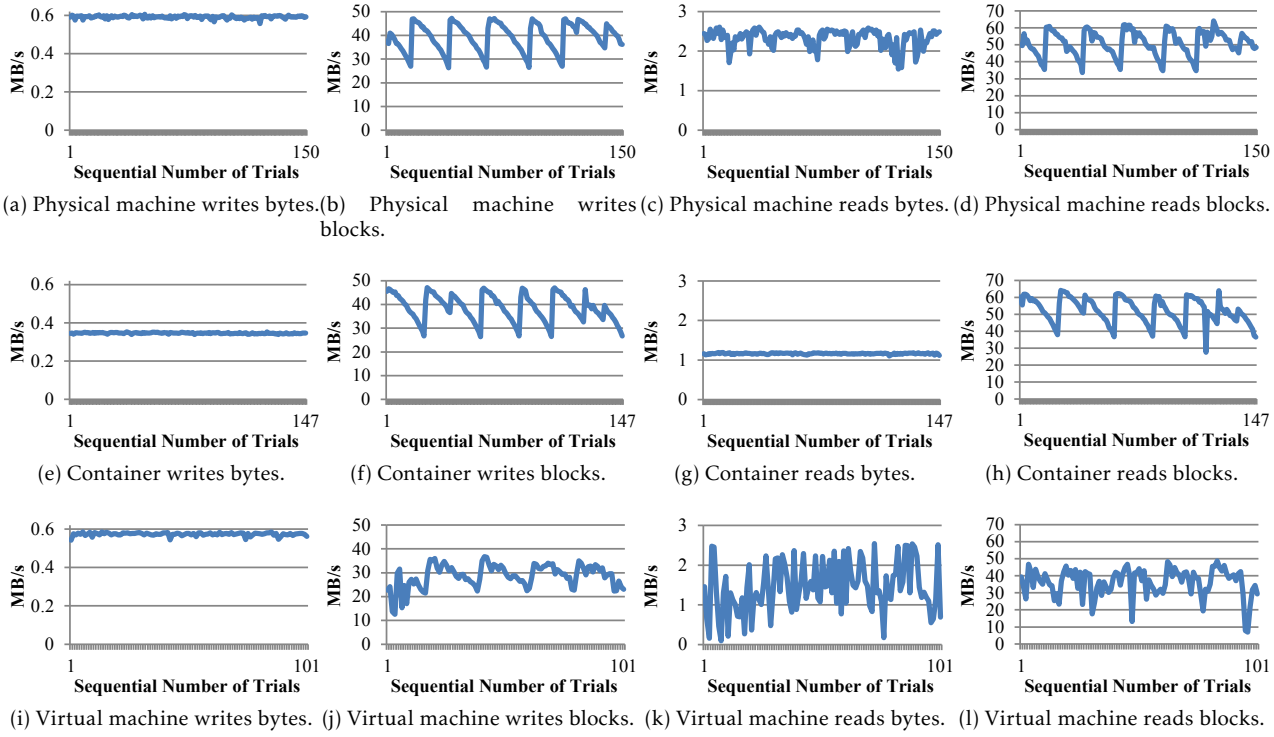


Figure 10: Storage data throughput and its variability overhead of a standalone Docker container vs. VM (using the benchmark Bonnie++).

3.4 Performance Evaluation Conclusion

Following the performance evaluation methodology DoKnowMe, we draw conclusions mainly by answering the predefined requirement questions. Driven by RQ1 and RQ2, our evaluation result largely confirms the aforementioned qualitative discussions: The container's average performance is generally better than the VM's and is even comparable to that of the physical machine with regarding to many features. Specifically, the container has less than 4% performance overhead in terms of communication data throughput, computation latency, memory data throughput and storage data throughput. Nevertheless, the container-based virtualization could hit a bottleneck of storage transaction speed, with the overhead up to 50%. Note that, as mentioned previously, we interpret the byte-size data throughput into storage transaction speed, because each byte essentially calls a disk transaction here. In contrast, although the VM delivers the worst performance in most cases, it could perform as well as the physical machine when solving the N-Queens problem or writing small-size data to the disk. By further comparing the storage filesystems of those two types of virtualization technologies, we believe that it is the copy-on-write mechanism that makes containers poor at storage transaction speed.

Driven by RQ3 and RQ4, we find that the performance loss resulting from virtualizations is more visible in the performance variability. For example, the container's variability overhead could reach as high as over 500% with respect to the Fibonacci calculation and the memory Triad operation. Similarly, although the container generally shows less performance variability than the VM, there are still exceptional cases: The container has the largest performance variation in the job of computing Fourier transform, whereas even the VM's performance variability is not worse than



(a) Physical machine writes bytes. (b) Physical machine writes blocks. (c) Physical machine reads bytes. (d) Physical machine reads blocks.

(e) Container writes bytes. (f) Container writes blocks. (g) Container reads bytes. (h) Container reads blocks.

(i) Virtual machine writes bytes. (j) Virtual machine writes blocks. (k) Virtual machine reads bytes. (l) Virtual machine reads blocks.

Figure 9: Storage benchmarking results by using Bonnie++ during 24 hours. The maximum x-axis scale indicates the iteration number of the Bonnie++ test (i.e. the physical machine, the container and the VM run 150, 147 and 101 tests respectively).

the physical machine's when running CryptoHash, N-Queens, and Raytracing jobs.

4 Container-based Application Case Study and Performance Optimization

4.1 Motive from a Background Project

Adequate pricing techniques play a key role in successful Cloud computing [20]. In the de facto Cloud market, there are generally three typical pricing schemes, namely on-demand pricing scheme, reserved pricing scheme, and spot pricing scheme. Although the fixed pricing schemes are dominant approaches to trading Cloud resources nowadays, spot pricing has been broadly agreed as a significant supplement for building a full-fledged market economy for the Cloud ecosystem [21]. Similar to the dynamic pricing in the electricity distribution industry, the spot pricing scheme here also employs a market-driven mechanism to provide spot service at a reduced and fluctuating price, in order to attract more demands and better utilize idle compute resources [22].

Unfortunately, the backend details behind changing spot prices are invisible for most of the Cloud market participants. In fact, unlike the static and straightforward pricing schemes of on-demand and reserved Cloud services, the market-driven mechanism for pricing Cloud spot service has been identified to be complicated both for providers to implement and for consumers to understand.

Therefore, it has become popular and valuable to take Amazon's spot service as a practical example to investigate Cloud spot pricing, so as to encourage and facilitate more players to enter the Cloud spot market. We are currently involved in a project on Cloud spot pricing analytics by using the whole-year price history of Amazon's 1053 types of spot service instances. Since Amazon only offers the most recent 60-day price trace to the public for review, we downloaded the price traces monthly in the past year to make sure the completeness of the whole-year price history.

In this way of downloading price traces, it is clear that around half of the overall raw data are duplicate. Moreover, the original price trace is sorted by the timestamp only, as shown in Table 6. To analyze spot service pricing on an instance-by-instance basis, however, the collected price data need not only to be sorted by timestamp but also to be distinguished by the other three attributes, i.e. Instance Type, Operating System and Zone. Thus, we decided to implement a preprocessing program to help clean the data, including removing the duplicate price records and categorizing price traces for individual service instances.

Tag	Price ($)	Timestamp	Instance Type	Operating System	Zone
SPOTINSTANCEPRICE	0.072700	2017-10-04T13:39:11+0000	m1.large	Windows	us-east-1a
SPOTINSTANCEPRICE	0.552500	2017-10-04T13:39:11+0000	c3.8xlarge	SUSE Linux	us-east-1b
SPOTINSTANCEPRICE	0.546700	2017-10-04T13:39:11+0000	c3.8xlarge	SUSE Linux	us-east-1e
SPOTINSTANCEPRICE	0.452500	2017-10-04T13:39:11+0000	c3.8xlarge	Linux/UNIX	us-east-1b
SPOTINSTANCEPRICE	0.446700	2017-10-04T13:39:11+0000	c3.8xlarge	Linux/UNIX	us-east-1e
SPOTINSTANCEPRICE	0.400800	2017-10-04T13:39:11+0000	c3.2xlarge	Windows (Amazon VPC)	us-east-1e
SPOTINSTANCEPRICE	0.417800	2017-10-04T13:39:11+0000	c3.4xlarge	SUSE Linux	us-east-1c
SPOTINSTANCEPRICE	0.317800	2017-10-04T13:39:11+0000	c3.4xlarge	Linux/UNIX	us-east-1c
SPOTINSTANCEPRICE	0.039600	2017-10-04T13:39:10+0000	m1.small	SUSE Linux	us-east-1b
SPOTINSTANCEPRICE	0.009600	2017-10-04T13:39:10+0000	m1.small	Linux/UNIX	us-east-1b
SPOTINSTANCEPRICE	0.434100	2017-10-04T13:39:10+0000	i3.2xlarge	SUSE Linux (Amazon VPC)	us-east-1c
SPOTINSTANCEPRICE	0.334100	2017-10-04T13:39:10+0000	i3.2xlarge	Linux/UNIX (Amazon VPC)	us-east-1c
SPOTINSTANCEPRICE	0.200000	2017-10-04T13:39:09+0000	c4.large	SUSE Linux (Amazon VPC)	us-east-1d
SPOTINSTANCEPRICE	0.100000	2017-10-04T13:39:09+0000	c4.large	Linux/UNIX (Amazon VPC)	us-east-1d
SPOTINSTANCEPRICE	0.281700	2017-10-04T13:39:09+0000	m3.2xlarge	Windows (Amazon VPC)	us-east-1c
SPOTINSTANCEPRICE	0.211200	2017-10-04T13:39:09+0000	m2.xlarge	SUSE Linux	us-east-1d
SPOTINSTANCEPRICE	0.111200	2017-10-04T13:39:09+0000	m2.xlarge	Linux/UNIX	us-east-1d

Table 6: A small piece of Amazon's spot price trace.

4.2 WordCount-alike Solution

Given the aforementioned requirement of data cleaning, we propose a WordCount-alike solution by analogy. WordCount is a well-known application that calculates the numbers of occurrences of different words within a document or word set. In the domain of big data analytics, WordCount is now a classic application to demonstrate the MapReduce mechanism that has become a standard programming model since 2004 [23]. As shown in Figure 11, the key steps in a MapReduce workflow are:

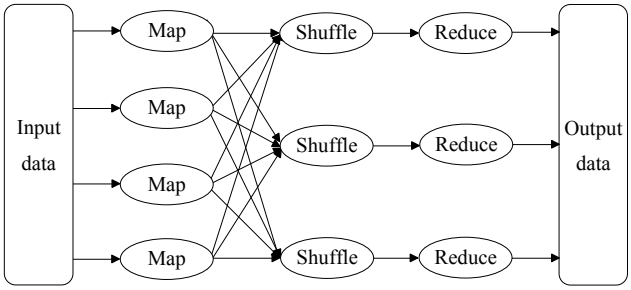


Figure 11: Workflow of the MapReduce process.

(1) The initial input source data are segmented into blocks according to the predefined split function and saved as a list of key-value pairs.

(2) The mapper executes the user-defined map function which generates intermediate key-value pairs.

(3) The intermediate key-value pairs generated by mapper nodes is sent to a specific reducer based on the key.

(4) Each reducer computes and reduces the data to one single key-value pair.

(5) All the reduced data are integrated into the final result of a MapReduce job.

Benefiting from MapReduce, applications like WordCount can deal with large amounts of data parallelly and distributedly. For the purpose of conciseness, we use a three-file scenario to demonstrate the process of MapReduce-based WordCount, as illustrated in Figure 12. In brief, the input files are broken into a set of <key, value> pairs for individual words, then the <key, value> pairs are shuffled alphabetically to facilitate summing up the values (i.e. the occurrence counts) for each unique key, and the reduced results are also a set of <key, value> pairs that directly act as the output in this case.

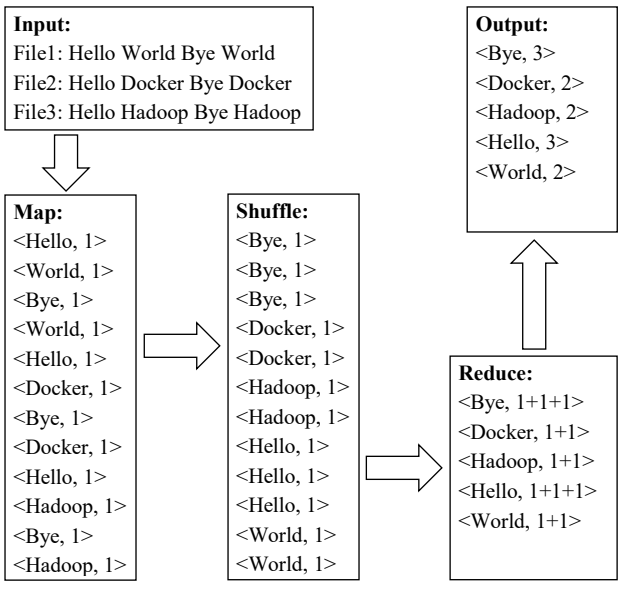


Figure 12: A Three-file Scenario of MapReduce-based WordCount.

Recall that the information of Amazon's spot price trace is composed of Tag, Price, Timestamp, Instance Type and Zone (cf. Table 6). By considering the value in each information field to be a letter, we treat every

single spot price record as an English word. For example, by random analogy, the spot price record "tag price1 time1 OS1 zone1" can be viewed as a six-letter word like "Hadoop", as highlighted in Figure 13.

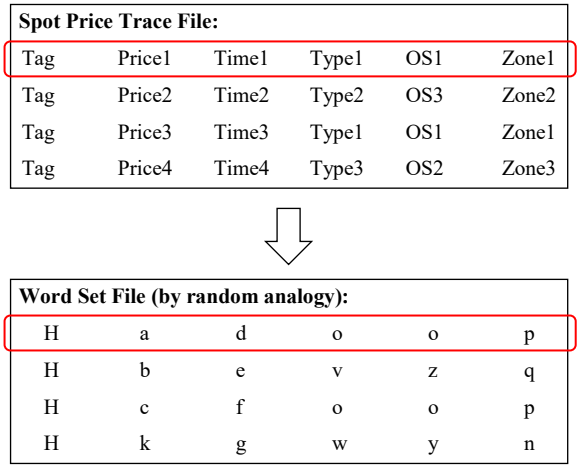


Figure 13: Treating spot price records as six-letter words.

As such, we are able to follow the logic of WordCount to fulfill the needs of cleaning our collected spot price history. In particular, instead of summing up the occurrences of the same price records, the duplicate records are simply ignored (removed). Moreover, the price records are sorted by the order of values of Instance Type, Operating System, Zone and Timestamp sequentially during the shuffling stage, while Tag and Price are not involved in data sorting. To avoid duplication, here we do not further elaborate the MapReduce-based data cleaning process.

4.3 Application Environment and Implementation

Unlike using "just-enough" environment for micro-level performance evaluation, here we employ "fair-enough" hardware resource to implement the MapReduce-based WordCount application, as listed in Table 7. In detail, the physical machine is Dell PowerEdge T110 II with the CPU model Intel Xeon E3-1200 series E3-1220 / 3.1 GHz. When preparing the MapReduce framework, we choose Apache Hadoop 2.7.0 [24] running in the operating system Ubuntu Server 16.04, and using Docker containers to construct a Hadoop cluster. In particular, Docker allows us to create an exclusive bridge network for the Hadoop cluster by using a specific name, e.g., `sudo docker network create --driver=bridge cluster`.

Environmental Item	Specification
Physical Machine	Dell PowerEdge T110 II
Operating System	Ubuntu Server 16.04
Java Environment	JDK 1.8.0
MapReduce Framework	Hadoop 2.7.0

Table 7: Summary of application environment.

In our initial implementation, we realized a three-node Hadoop cluster by packing Hadoop 2.7.0 into a Docker image and starting one master container and two slave ones. As illustrated in Figure 14, such a Hadoop cluster can logically be divided into a distributed processing layer (i.e. MapReduce workflow) and a distributed file-system layer (i.e. Hadoop Distributed File System (HDFS) in this case). When running a MapReduce application, the main interactions are:

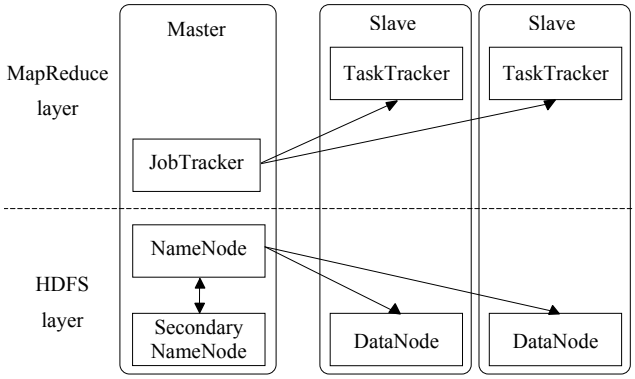


Figure 14: A three-node Hadoop cluster.

(1) A MapReduce job can run in a Hadoop cluster.

(2) The JobTracker in the cluster accepts a job from the MapReduce application, and locates relevant data through the NameNode.

(3) Suitable TaskTrackers are selected and then accept the tasks delivered by the JobTracker.

(4) The JobTracker communicates with the TaskTrackers and manages failures.

(5) When the collaboration among the TaskTrackers finish the job, the JobTracker updates its status and returns the result.

4.4 Performance Optimization Strategies

It is known that the performance of MapReduce applications can be tuned by adjusting the various parameters in the three configuration files of Hadoop. However, it is also clear that there is no one-size-fits-all approach to performance tuning. Therefore, we conducted a set of performance evaluation of our data cleaning application, in order to come up with a set of optimization strategies at least for this case.

- **Setting Timeout for Tasks.** A map or reduce task can be blocked or failed during runtime for various reasons, which would slow down the execution, and even result in the failure, of the whole MapReduce job. Therefore, by using TaskTracker to kill the blocked/failed tasks after a proper time span, those tasks will be able to be relaunched to save some waiting time. Given our relatively small size of data, we reduce the default timeout value from ten minutes to one minute, as shown below.

```
<!--Configuration in mapred-site.xml-->
<property>
  <name>mapred.task.timeout</name>
  <value>60000</value>
</property>
```

- **Turn on Out-of-Band Heartbeat.** Unlike regular heartbeats, the out-of-band heartbeat is triggered when a task is complete or failed. As such, JobTracker will be noticed the first time when there are free resources, so as to assign them to new tasks and eventually to save time. The configuration for turning on out-of-band heartbeat is specified as follows.

```
<!--Configuration in mapred-site.xml-->
<property>
  <name>mapreduce.tasktracker.
          outofband.heartbeat</name>
  <value>true</value>
</property>
```

- **Setting Buffer.** To begin with, we are concerned with a threshold percentage of buffer, and a background thread will be issued to spill buffer contents to hard disk when the threshold is reached. Inspired by the storage micro-benchmarking results (cf. Section 3.3.4), we decided to increase the threshold from 80% to 90% of buffer. As for the amount of memory to be buffer size, we double the default value (i.e. 100MB) for an intuitive test. These two parameters can be adjusted respectively as shown below.

```
<!--Configuration in mapred-site.xml-->
<property>
  <name>io.sort.spill.percent</name>
  <value>0.9</value>
</property>
<property>
  <name>io.sort.mb</name>
  <value>200</value>
</property>
```

- **Merging Spilled Streams.** As a continuation of spilling buffer contents to hard disk, the intermediate streams from multiple spill threads are merged into one single sorted file per partition which is to be fetched by reducers. Thus, we can control how many of spills will be merged into one file at a time. Since the smaller merge factor incurs more parallel merge activities and more disk IO for reducers, we decided to increase the merge value from 10 to 100, as shown below.

```
<!--Configuration in mapred-site.xml-->
<property>
  <name>io.sort.factor</name>
  <value>100</value>
</property>
```

- **LZO Compression.** Recall that the data we are dealing with are plain texts. The text data can generally be compressed significantly to reduce the usage of hard disk space and transmission bandwidth, and correspondingly to save the time taken for data copying/transferring. As a lossless algorithm with high decompression speed, Lempel-Ziv-Oberhumer (LZO) is one of the compression mechanisms supported by the Hadoop framework. In addition to various benefits and characteristics in common, LZO's block structure is particularly split-friendly for parallel processing in MapReduce jobs [25]. Therefore, we install and enable LZO for Hadoop by specifying the configurations below.

```
<!--Configuration in core-site.xml-->
<property>
  <name>io.compression.codecs</name>
  <value>org.apache.hadoop.io.compress.
          GzipCodec,org.apache.hadoop.
          io.compress.DefaultCodec,
       com.hadoop.compression.lzo.
          LzoCodec,com.hadoop.
          compression.lzo.LzopCodec,
       org.apache.hadoop.io.compress.
          BZip2Codec</value>
</property>
<property>
  <name>io.compression.codec.
                  lzo.class</name>
  <value>com.hadoop.compression.
              lzo.LzoCodec</value>
</property>

<!--Configuration in mapred-site.xml-->
<property>
  <name>mapreduce.map.output.
                  compress</name>
  <value>true</value>
</property>
<property>
  <name>mapreduce.map.output.
              compress.codec</name>
  <value>com.hadoop.compression.
               lzo.LzoCodec</value>
</property>
```

- **Doubling Slave Nodes.** In distributed computing, a common scenario is to employ more resources to deal with more workloads. Recall that map and reduce tasks of a MapReduce job are distributed to the slave nodes in a Hadoop cluster, and the physical machine used in this study has a Quad-Core processor. To obtain some quick clues at this current stage, we try to improve our application's performance by doubling the slave nodes, i.e. extending the original three-node cluster (with two slave nodes) into a five-node one (with four slave nodes). Note that, in this optimization strategy, we keep the other configurations settings by default.

4.5 Performance Evaluation Results

Due to the time limit, we follow "one factor at a time" to perform evaluation of the aforementioned optimization strategies. Furthermore, to confirm the effects of

these optimization strategies, we are concerned with 2GB+ and 5GB+ data respectively in the performance evaluation. Correspondingly, we draw the evaluation results in Figure 15 and 16 respectively.

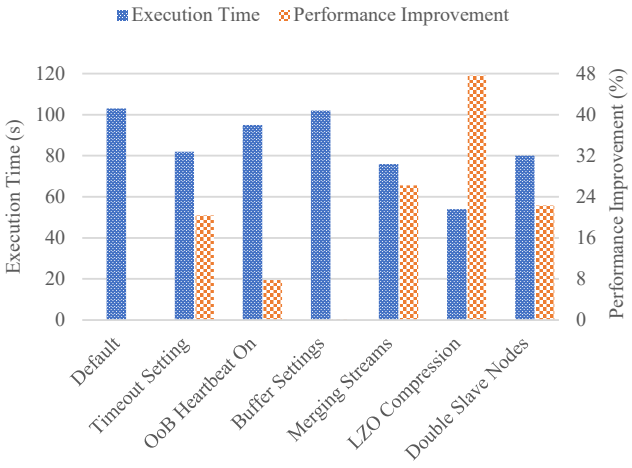


Figure 15: Performance optimization for 2GB+ data cleaning.

The results show three relatively different patterns of optimization effects. First, LZO compression acts as the most effective optimization strategy, and it can improve the performance nearly by 50% when dealing with 2GB+ data. Second, we can expect similar and moderate performance improvement by three individual strategies, such as merging more spilled streams, reducing the timeout value, and doubling the slave nodes. Third, out-of-band heartbeat and buffer settings seem not to be influential optimization strategies in this case.

These different optimization patterns could be closely related to the data characteristics of our application. On one hand, since text data can be compressed significantly [26], our application mostly benefits from the optimization strategy of LZO compression. On the other hand, since the current price traces used in this study are still far from "big data", the buffer setting could not take clear effects until the data size reaches TB levels.

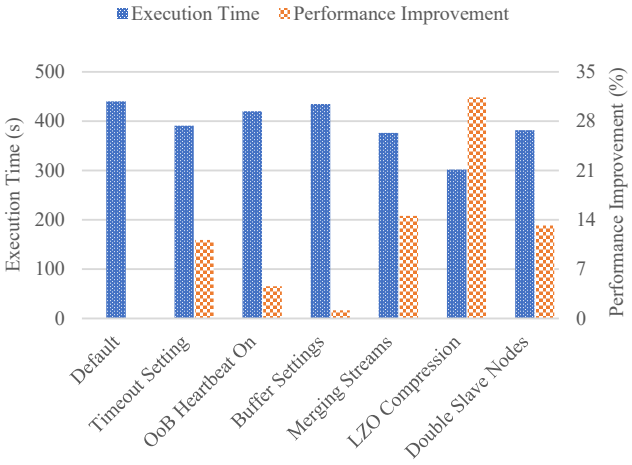


Figure 16: Performance optimization for 5GB+ data cleaning.

5 Related Work

Although the performance advantage of containers were investigated in several pioneer studies [3, 27, 10], the container-based virtualization solution did not gain significant popularity until the recent underlying improvements in the Linux kernel, and especially until the emergence of Docker [28, 29]. Starting from an open-source project in early 2013 [7], Docker quickly becomes the most popular container solution [2] by significantly facilitating the management of containers. Technically, through offering the unified tool set and API, Docker relieves the complexity of utilizing the relevant kernel-level techniques including the LXC, the cgroup and a copy-on-write filesystem. To examine the performance of Docker containers, a molecular modeling simulation software [30] and a postgreSQL database-based Joomla application [31] have been used to benchmark the Docker environment against the VM environment.

Considering the uncertainty of use cases (e.g., different workload densities and QoS requirements), at this current stage, a baseline-level investigation would be more useful and helpful for understanding the fundamental difference in performance overhead between those two virtualization solutions. A preliminary study has particularly focused on the CPU consumption by using the 100000! calculation within a Docker container and a KVM VM respectively [32]. Nevertheless, the concerns about other features/resources like memory and disk are missing. Similarly, the performance analysis between VM and container in study [33] is not feature-specific enough (even including security that is out of the scope of performance). On the contrary, by treating Docker containers as a particular type of Cloud service, our study considers the four physical properties of a Cloud service [15] and essentially gives a fundamental investigation into the Docker container's performance overhead on a feature-by-feature basis.

The closest work to ours is the IBM research report on the performance comparison of VM and Linux containers [34]. In fact, it is this incomplete report (e.g., the container's network evaluation is partially missing) that inspires our study. Surprisingly, our work denies the IBM report's finding "containers and VMs impose almost no overhead on CPU and memory usage" that was also claimed in [35], and we also doubt about "Docker equals or exceeds KVM performance in every case". In particular, we are more concerned with the overhead in performance variability.

Within the context of MapReduce clusters, Xavier et al. [10] conducted experimental comparisons among the three aforementioned types of container-based virtual environments, while a set of other studies particularly contrasted performance of OpenVZ with the hypervisor-based virtualization implementations including VMWare, Xen and KVM [3, 27]. The significant difference between these studies and ours is that we focus more on the performance optimization of a

container-based MapReduce cluster in a specific application scenario.

Note that, although there are also performance studies on deploying containers inside VMs (e.g., [36, 37]), such a redundant structure might not be suitable for an "apple-to-apple" comparison between Docker containers and VMs, and thus we do not include this virtualization scenario in our study.

6 Conclusions and Future Work

It has been identified that virtualization is one of the foundational elements of Cloud computing and helps realize the value of Cloud computing [38]. On the other hand, the technologies for virtualizing Cloud infrastructures are not resource-free, and their performance overheads would incur negative impacts on the QoS of the Cloud. Since hypervisors that currently dominate the Cloud virtualization market are a relatively heavyweight solution, there comes a rising trend of interest in its lightweight alternative [7], namely the container-based virtualization. Their mechanism difference is that, the former manages the host hardware resources, while the latter enables sharing the host OS. Although straightforward comparisons can be done from the existing qualitative discussions, we conducted a fundamental evaluation study to quantitatively understand the performance overheads of these two different virtualization solutions. In particular, we employed a standalone Docker container and a VMWare Workstation VM to represent the container-based and the hypervisor-based virtualization technologies respectively.

Recall that there are generally two stages of performance engineering in ECS, for revealing the primary performance of specific (system) features and investigating the overall performance of real-world applications respectively. In addition to the fundamental performance of a single container, we also studied performance optimization of a container-based MapReduce application in terms of cleaning Amazon's spot price history. At this current stage, we only focused on one factor at a time to evaluate the optimization strategies ranging from setting task timeout to doubling slave nodes.

Overall, our work reveals that the performance overheads of these two virtualization technologies could vary not only on a feature-by-feature basis but also on a job-to-job basis. Although the container-based solution is undoubtedly lightweight, the hypervisor-based technology does not come with higher performance overhead in every case. At the application level, the container technology is clearly more resource-friendly, as we failed in building VM-based MapReduce clusters on the same physical machine. When it comes to container-based MapReduce applications, it seems that the effects of performance optimization strategies are closely related to the data characteristics. For dealing with text data in our case study, LZO compression can bring the most significant performance improvement.

Due to the time and resource limit, our current investigation into the performance of container-based MapReduce applications is still an early study. Thus, our future work will be unfolded along two directions. Firstly, we will adopt sophisticated experimental design techniques (e.g., the full-factorial design) [39] to finalize the same case study on tuning the MapReduce performance of cleaning Amazon's price history. Secondly, we will gradually apply Docker containers to different real-world applications for dealing with different types of data. By employing "more-than-enough" computing resource, the application-oriented practices will also be replicated in the hypervisor-based virtual environment for further comparison case studies.

Acknowledgment This work is supported by the Swedish Research Council (VR) under contract number C0590801 for the project Cloud Control, and through the LCCC Linnaeus and ELLIIT Excellence Centers. This work is also supported by the National Natural Science Foundation of China (Grant No.61572251).

References

[1] Z. Li, M. Kihl, Q. Lu, and J. A. Andersson, "Performance overhead comparison between hypervisor and container based virtualization," in *Proceedings of the 31st IEEE International Conference on Advanced Information Networking and Application (AINA 2017)*. Taipei, Taiwan: IEEE Computer Society, 27-29 March 2017, pp. 955–962, https://doi.org/10.1109/AINA.2017.79.

[2] C. Pahl, "Containerization and the PaaS Cloud," *IEEE Cloud Computing*, vol. 2, no. 3, pp. 24–31, May/June 2014, https://doi.org/10.1109/MCC.2015.51.

[3] J. P. Walters, V. Chaudhary, M. Cha, S. G. Jr., and S. Gallo, "A comparison of virtualization technologies for HPC," in *Proceedings of the 22nd International Conference on Advanced Information Networking and Applications (AINA 2008)*. Okinawa, Japan: IEEE Computer Society, 25-28 March 2008, pp. 861–868, https://doi.org/10.1109/AINA.2008.45.

[4] P. J. Denning, "Performance evaluation: Experimental computer science at its best," *ACM SIGMETRICS Performance Evaluation Review*, vol. 10, no. 3, pp. 106–109, Fall 1981, https://doi.org/10.1145/1010629.805480.

[5] D. G. Feitelson, "Experimental computer science," *Communications of the ACM*, vol. 50, no. 11, pp. 24–26, November 2007, https://doi.org/10.1145/1297797.1297817.

[6] Z. Li, L. O'Brien, and M. Kihl, "DoKnowMe: Towards a domain knowledge-driven methodology for performance evaluation," *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, no. 4, pp. 23–32, 2016, https://doi.org/10.1145/2897356.2897360.

[7] D. Merkel, "Docker: Lightweight Linux containers for consistent development and deployment," *Linux Journal*, vol. 239, pp. 76–91, March 2014.

[8] X. Xu, H. Yu, and X. Pei, "A novel resource scheduling approach in container based clouds," in *Proceedings of the 17th IEEE International Conference on Computational Science and Engineering (CSE 2014)*. Chengdu, China: IEEE Computer Society, 19-21 December 2014, pp. 257–264, https://doi.org/10.1109/CSE.2014.77.

[9] D. Bernstein, "Containers and Cloud: From LXC to Docker to KuBernetes," *IEEE Cloud Computing*, vol. 1, no. 3, pp. 81–84, September 2014, https://doi.org/10.1109/MCC.2014.51.

[10] M. G. Xavier, M. V. Neves, and C. A. F. D. Rose, "A performance comparison of container-based virtualization systems for MapReduce clusters," in *Proceedings of the 22nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP 2014)*. Turin, Italy: IEEE Press, 12-14 February 2014, pp. 299–306, https://doi.org/10.1109/PDP.2014.78.

[11] C. Anderson, "Docker," *IEEE Software*, vol. 32, no. 3, pp. 102–105, May/June 2015, https://doi.org/10.1109/MS.2015.62.

[12] T. Banerjee, "Understanding the key differences between LXC and Docker," https://www.flockport.com/lxc-vs-docker/, August 2014.

[13] S. M. Blackburn, K. S. McKinley, R. Garner, C. Hoffmann, A. M. Khan, R. Bentzur, A. Diwan, D. Feinberg, D. Frampton, S. Z. G. M. H. A. H. M. J. H. Lee, J. E. B. Moss, A. Phansalkar, D. Stefanovik, T. VanDrunen, D. von Dincklage, and B. Wiedermann, "Wake up and smell the coffee: Evaluation methodology for the 21st century," *Communications of the ACM*, vol. 51, no. 8, pp. 83–89, August 2008, https://doi.org/10.1145/1378704.1378723.

[14] A. Iosup, N. Yigitbasi, and D. Epema, "On the performance variability of production Cloud services," in *Proceedings of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2011)*. Newport Beach, CA, USA: IEEE Computer Society, 23-26 May 2011, pp. 104–113, https://doi.org/10.1109/CCGrid.2011.22.

[15] Z. Li, L. O'Brien, R. Cai, and H. Zhang, "Towards a taxonomy of performance evaluation of commercial Cloud services," in *Proceedings of the 5th International Conference on Cloud Computing (IEEE CLOUD 2012)*. Honolulu, Hawaii, USA: IEEE Computer Society, 24-29 June 2012, pp. 344–351, https://doi.org/10.1109/CLOUD.2012.74.

[16] Z. Li, L. O'Brien, H. Zhang, and R. Cai, "A factor framework for experimental design for performance evaluation of commercial Cloud services," in *Proceedings of the 4th International Conference on Cloud Computing Technology and Science (CloudCom 2012)*. Taipei, Taiwan: IEEE Computer Society, 3-6 December 2012, pp. 169–176, https://doi.org/10.1109/CloudCom.2012.6427525.

[17] StackOverflow, "What is the relationship between the docker host OS and the container base image OS?" http://stackoverflow.com/questions/18786209/what-is-the-relationship-between-the-docker-host-os-and-the-container-base-image, September 2013.

[18] Reddit, "Do I need to use an OS base image in my Dockerfile or will it default to the host OS?" https://www.reddit.com/r/docker/comments/2teskf/do_i_need_to_use_an_os_base_image_in_my/, January 2015.

[19] Z. Li, L. O'Brien, H. Zhang, and R. Cai, "On the conceptualization of performance evaluation of IaaS services," *IEEE Transactions on Services Computing*, vol. 7, no. 4, pp. 628–641, October-December 2014, https://doi.org/10.1109/TSC.2013.39.

[20] C. Weinhardt, A. Anandasivam, B. Blau, N. Borissov, T. Meinl, W. Michalk, and J. Stößer, "Cloud computing a classification, business models, and research directions," *Business and Information Systems Engineering*, vol. 1, no. 5, pp. 391–399, October 2009, https://doi.org/10.1007/s12599-009-0071-2.

[21] V. Abhishek, I. A. Kash, and P. Key, "Fixed and market pricing for Cloud services," in *Proceedings of the 7th Workshop on the Economics of Networks, Systems, and Computation (NetEcon 2012)*. Orlando, FL, USA: IEEE Computer Society, 20 March 2012, pp. 157–162, https://doi.org/10.1109/INFCOMW.2012.6193479.

[22] Z. Li, H. Zhang, L. O'Brien, S. Jiang, Y. Zhou, M. Kihl, and R. Ranjan, "Spot pricing in the cloud ecosystem: A comparative investigation," *Journal of Systems and Software*, vol. 114, pp. 1–19, April 2016, https://doi.org/10.1016/j.jss.2015.10.042.

[23] Q. Lu, Z. Li, M. Kihl, L. Zhu, and W. Zhang, "CF4BDA: A conceptual framework for big data analytics applications in the cloud," *IEEE Access*, vol. 3, pp. 1944–1952, October 2015, https://doi.org/10.1109/ACCESS.2015.2490085.

[24] Apache Software Foundation, "Mapreduce tutorial," https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html, October 2017.

[25] A. Viswanathan, "A guide to using LZO compression in Hadoop," *Linux Journal*, vol. 2012, no. 220, August 2012, article no. 1.

[26] WinZip, "Varying file compression explored," http://kb.winzip.com/kb/entry/326/, September 2017.

[27] J. Che, C. Shi, Y. Yu, and W. Lin, "A synthetical performance evaluation of OpenVZ, Xen and KVM," in *Proceedings of the 2010 IEEE Asia-Pacific Services Computing Conference (APSCC 2010)*. Hangzhou, China: IEEE Computer Society, 6-10 December 2010, pp. 587–594, https://doi.org/10.1109/APSCC.2010.83.

[28] D. Strauss, "Containers - not virtual machines - are the future Cloud," *Linux Journal*, vol. 228, pp. 118–123, April 2013.

[29] A. Karle, "Operating system containers vs. application containers," https://blog.risingstack.com/operating-system-containers-vs-application-containers/, May 2015.

[30] T. Adufu, J. Choi, and Y. Kim, "Is container-based technology a winner for high performance scientific applications?" in *Proceedings of the 17th Asia-Pacific Network Operations and Management Symposium (APNOMS 2015)*. Busan, Korea: IEEE Press, 19-21 August 2015, pp. 507–510, https://doi.org/10.1109/APNOMS.2015.7275379.

[31] A. M. Joy, "Performance comparison between Linux containers and virtual machines," in *Proceedings of the 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA 2015)*. Ghaziabad, India: IEEE Press, 14-15 February 2015, pp. 507–510, https://doi.org/10.1109/ICACEA.2015.7164727.

[32] K.-T. Seo, H.-S. Hwang, I.-Y. Moon, O.-Y. Kwon, and B.-J. Kim, "Performance comparison analysis of Linux container and virtual machine for building Cloud," *Advanced Science and Technology Letters*, vol. 66, pp. 105–111, December 2014, https://doi.org/10.14257/astl.2014.66.25.

[33] R. K. Barik, R. K. Lenka, K. R. Rao, and D. Ghose, "Performance analysis of virtual machines and containers in cloud computing," in *Proceedings of the 2016 International Conference on Computing, Communication and Automation (ICCCA 2016)*. Greater Noida, India: IEEE Press, 29-30 April 2016, pp. 1204–1210, https://10.1109/CCAA.2016.7813925.

[34] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio, "An updated performance comparison of virtual machines and Linux containers," in *Proceedings of the 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS 2015)*. Philadelphia, PA, USA: IEEE Press, 29-31 March 2015, pp. 171–172, https://doi.org/10.1109/ISPASS.2015.7095802.

[35] Z. Kozhirbayev and R. O. Sinnott, "A performance comparison of container-based technologies for the Cloud," *Future Generation Computer Systems*, no. 68, pp. 175–182, March 2017, https://doi.org/10.1016/j.future.2016.08.025.

[36] R. Dua, A. R. Raja, and D. Kakadia, "Virtualization vs containerization to support PaaS," in *Proceedings of the 2014 IEEE International Conference on Cloud Engineering (IC2E 2015)*. Boston, Massachusetts, USA: IEEE Computer Society, 10-14 March 2014, pp. 610–614, https://doi.org/10.1109/IC2E.2014.41.

[37] S. F. Piraghaj, A. V. Dastjerdi, R. N. Calheiros, and R. Buyya, "Efficient virtual machine sizing for hosting containers as a service," in *Proceedings of the 11th World Congress on Services (SERVICES 2015)*. New York, USA: IEEE Computer Society, 27 June - 2 July 2015, pp. 31–38, https://doi.org/10.1109/SERVICES.2015.14.

[38] S. Angeles, "Virtualization vs. Cloud computing: What's the difference?" http://www.businessnewsdaily.com/5791-virtualization-vs-cloud-computing.html, January 2014.

[39] D. C. Montgomery, *Design and Analysis of Experiments*, 8th ed. Hoboken, NJ: John Wiley & Sons, Inc., April 2012.