ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Enhanced emotion recognition by feature selection to animate a talking head

Hela DAASSI-GNABA[1] and Yacine OUSSAR[2]

1- Université Paris 8 - Laboratoire d'Informatique Avancée de Saint-Denis (LIASD, EA 4383) 2 rue de la Liberté, 93526 Saint-Denis Cedex - France

2- ESPCI-ParisTech - Laboratoire de Physique et d'Étude des Matériaux (LPEM) 10 rue Vauquelin, 75231 Paris Cedex - France

Abstract. It is known that deaf and hard of hearing people can substantially improve their skill to lip reading if they have access to speaker emotion. Moreover, it has been shown that animating an artificial talking head can provide this modality. In this paper, we assume that emotion recognition to animate such talking head can be performed using a small set of relevant features extracted from the speech signal. More precisely, we show that the implementation of linear classifiers using Support Vector Machines (SVM) with the involvement of a feature selection method leads to a promising performance which confirms our assumption.

1 Introduction

A direct speech to facial animation conversion system was recently developed as a communication aid for deaf and hard of hearing people [1]. These people have great abilities in understanding speech based on lip reading only. To help them improving such skill, previous studies proposed to provide to these people an artificial talking head controlled by automatic speech recognition [2]. However, results showed that many users need more than that. Indeed, they need the emotional state of the speaker in order to fully understand a given speech.

The present study proposes to enhance the talking head by adding the emotion modality using acoustic information extracted from speech [3]. The acoustic information is given by a set of both prosodic and spectral features. These features are ranked and selected according to their level of relevance in order to best optimize the emotion recognition performed by classification. Thus, the classifier transmits the emotions to the talking head so it can be animated.

The resulting system can be implemented in various domains as a key factor to improve the quality of life of deaf and hard of hearing people. For instance, the current device might helps in: school, education institutions, meetings, conferences and telecommunication systems that become more accessible to them.

In addition, the speech signal compiled to animate the talking head using the proposed approach is informative on both the speech content and its prosody. In practice, the talking head can be displayed on a PDA (Personal Digital Assistant) or a computer screen placed face to the deaf person [4].

The paper is organized as follows: section 2 describes the approach we propose and some implementation issues. Section 3 presents the corpus used for

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

emotion detection, the acoustic features used and the extraction process. Finally, section 4 illustrates the results obtained by numerical experiments for emotion recognition.

2 From the speech signal to the talking head

In this section, we describe our approach which can be summarized by the block diagram illustrated on Figure 1.

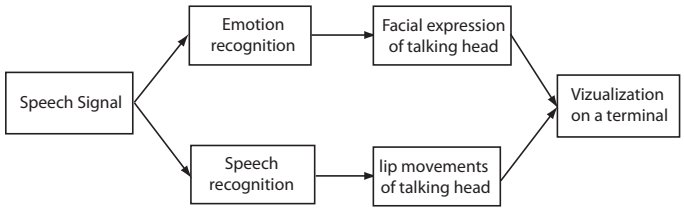

Fig. 1: Block diagram of the combined system.

It consists in animating a talking head using speech signal. Since emotions represent an important modality when communicating a message, the use of emotions to animate a talking head can contribute to various domains of life.

- Speech recognition: For our application, we used Dragon Naturally Speaking in Automatic Speech Recognition (ASR). The software turns speech to text at up to 160 words-per-minute (version 11). The spoken words, which are recorded using a microphone and converted into a text, are then processed by a special translator into lip reading.
- Emotion recognition: The emotion detection system is based on a SVM linear classifier that is trained on the LDC (Linguistic Data Consortium) corpus [5] (see Section 4). The decision of the classifier is the emotional state of the speaker: this piece of information is used to animate the facial expression of the talking head.

Fig. 2: Greta with happy (left) and neutral (right) facial expressions.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

- The talking head Greta: The Greta model proposed in [6] is able to communicate using a rich palette of verbal and nonverbal behaviors. The agent can talk and simultaneously show facial expressions, gestures, gaze, and head movements. The speech is provided to the Greta system as a text file. In addition, communicative intentions and behaviors of Greta can also be added into the text file using APML (Affective Presentation Markup Language). Greta can speak various languages. In this work Greta speaks English since our data corpus is in the English language.
  In this study, we focus on recognizing two facial expressions that correspond to happy (Figure 2-left) and neutral emotions (Figure 2-right).

3 Corpus and acoustic feature extraction

The data used in our experiments were obtained from the LDC Emotional Prosody and Transcripts database [5]. It consists in English language acted speech recordings. This database contains both audio recordings and the corresponding transcripts. The recordings deal with professional actors reading series of semantically neutral utterances (dates and numbers) spanning fifteen distinct emotional categories. In this study, we chose six professional actors: three males (CC, CL and MF) and three females (GG, JG and MK) to guarantee speaker gender balancing. From the LDC Emotional Prosody Data, we focused on the recognition of "happy" versus "neutral" emotions.
Pre-processing the database leads to 53 utterances for female speakers (32 happy and 21 neutral utterances) and 52 for male (31 happy and 21 neutral utterances).

Feature extraction is a fundamental issue when dealing with data separation. In one hand, the candidate features must form a large set in order to be enough informative and selective to separate the data. In the other, only the most relevant of them have to be involved in the design of the classifier. Irrelevant features can be considered as noisy data and may deteriorate the classification accuracy. In order to take into account relevant and informative input data, we considered both prosodic and spectral features.
More specifically, we propose to use statistical moments from two prosodic features and four spectral features. The statistical moments are : maximum, minimum, range (maximum-minimum), mean, median and standard deviation. The selected prosodic features are the fundamental frequency contour (F_0) and the energy contour (E_n). The selected spectral features are the formant frequencies (F_1, F_2) and their bandwidths (B_1, B_2). We gathered the 6 moments from each of the 6 acoustic features to generate a set of $N_{max} = 36$ variables. All these features were extracted with the Praat program [7] and their statistical moments were computed using the Matlab software. According to the corpus described above, 105 utterances are available to separate the happy versus neutral emotions. We assume that each utterance is fully described by providing values for at most the 36 available features.

The Gram-Schmidt orthogonalization procedure [8] was implemented to rank the 36 candidate features according to their relevance. Once the features are

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

ranked, the most relevant of them can be selected using either a filter or a wrapper approach [9]. Although the filter approach is often computationally cost effective, we focused on the wrapper approach which usually leads to a better generalization in the data separation with an acceptable computational burden in our implementation.

4 Classification and experimental results

4.1 Support Vector Machines (SVM)

SVM classification is used to find an optimal separation between two classes : the maximum margin hyperplane [10]. The set of examples that are sufficient to determine the maximum margin hyperplane are called the support vectors. If the data are linearly separable, a linear SVM classifier is sufficient. Otherwise, if the data are not linearly separable, SVM classification proceeds by projecting the input vectors in high dimensional space called the feature space then a linear separation is possible. In practice, this data conversion leads to the use of a kernel function. To be a SVM kernel, a function has to verify a set of conditions listed in [10]. An SVM discriminant function is given by:

$$f(x) = \sum_{i=1}^{M} \alpha_i y_i k(x, x_i) + b \tag{1}$$

where: k is the kernel function, x_i are the support vectors, y_i are the corresponding class labels (± 1) and M is the number of support vectors. Note that $\alpha_i$and b are the parameters of the classifier adjusted during the training process. A regularization parameter C controls the trade off between classification errors on training data and margin maximization.
The validation procedure implemented in our experiments is a special case of the cross validation method. It is called the Leave-One-Speaker-Out (LOSO) method. This procedure consists in a data partitioning where the validation fold contains data from the sixth speaker that does not appear in the training folds containing data from the five other speakers. This method guarantees strict speaker independence.

4.2 Experimental results

The Ho-Kashyap algorithm [11] was first run to determine if the data are linearly separable regarding the training examples.
When taking into account the whole available examples (105 utterances), the Ho-Kashyap algorithm showed that the data are linearly separable when the $N = 15$ most relevant input variables over the 36 available are used. Hence, linear SVM are preferred at first sight and implemented using the spider software [12]. Note that the discriminant function of a linear SVM is given by equation 1 with the identity function as a kernel. Since the Gram-Schmidt selection procedure was implemented according to a wrapper scheme, both the regularization parameter

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

C and the number of the selected relevant variables were jointly optimized using the 6-fold LOSO validation procedure.
The best validation score was obtained with a linear SVM having $N = 15$ and $C = 20$: 89.5% of the emotions were successfully recognized. As mentioned above, such linear SVM with the 15 most relevant features as inputs ensures data linear separability when all the available examples are involved in the training set. Since we implemented the LOSO validation procedure, the behavior of the validation speaker can be slightly different from the others. Hence, the hyperplane determined during training may not successfully separate all the examples that belong to the validation speaker. As a result, the overall recognition rate is not guaranteed to be close to 100%.
Table 1 summarizes the results of the classification of happy versus neutral emotions for each speaker (first column); we show the percentage of well classified items (second column), the precision (third column) and the recall for the recognition of happy and neutral emotions (fourth column). For a given class, the precision is the fraction of well classified emotions over all those put in this class. The recall is the fraction of emotions put in this class over all those labeled in this class. Precision and recall are computed using the confusion matrices.

Speaker		Well classified	Precision		Recall	
			Happy	Neutral	Happy	Neutral
Male	CC	77.8%	0.89	0.67	0.73	0.86
	CL	100%	1	1	1	1
	MF	100%	1	1	1	1
Female	GG	88.9%	0.91	0.86	0.91	0.86
	JG	83.3%	0.83	0.83	0.91	0.71
	MK	88.2%	1	0.78	0.8	1

Table 1: Percentage of well classified emotions, precision and recall.

These results show that a promising rate in emotion recognition can be obtained with few relevant acoustic features. This classification confirms the feasibility of our approach as well as a numerically cost effective implementation to animate the talking head. Indeed, the classifier is linear and uses a small set of input variables instead of the overall 36 extracted from the speech signal.
As a comparison, the percentage of well classified emotions obtained with the same corpus when using the Weka software [13] and the 36 available features is 74.3%. We assume that this rate is worser since Weka does not implement any efficient feature selection method and neglects the optimization of parameter C.
For the sake of the design of parsimonious classifiers involving less input variables ($N < 15$), nonlinear SVM classifiers using a Gaussian kernel were also implemented. The regularization parameter C, the Gaussian kernel parameter σ, as well as the value of N were simultaneously optimized according to the 6-fold LOSO validation procedure described above. The results showed that neither the classification error was decreased nor the number of relevant input variables was optimized. Hence, we consider that the non linear classification does not

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

bring any improvement for the "happy" versus "neutral" emotions separation.

5 Conclusion and perspectives

To improve the independence of deaf and hard of hearing people in their everyday life, a system based on emotion recognition using acoustic information extracted from speech to animate a talking head is described in this paper. We mainly focused on the optimization of the emotion recognition method. Numerical experiments conducted using SVM classification involving feature ranking and selection showed that a promising rate of recognition can be obtained with a linear classifier involving a small set of relevant input variables ensuring even data linear separability.
Future work will focus on introducing more different emotions in order to improve the overall sensitivity. Moreover, further refinements will include new features related to the emotional state of the speaker particularly by paying attention to the selection of both acoustic and linguistic descriptors to best optimize the design of multi-class classifiers.

References

[1] G. Takćs, A. Tihanyi, T. Bárdi, G. Feldhoffer and B. Srancsik, Speech to facial animation conversion for deaf customers, *proceedings of the* 14^{th} *European Signal Processing Conference* (EUSIPCO 2006), September 4-8, Florence (Italy), 2006.

[2] Labiao. Available online: http://labiao.it-sudparis.eu/

[3] M. El Ayadi, M. S. Kamel and F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition*, 44:572-587, 2011.

[4] H. Daassi-Gnaba and J. Lopez Krahe, Universal combined system: speech recognition, emotion recognition and talking head for deaf and hard of hearing people, *proceedings of the* 10^{th} *Association for the Advancement of Assistive Technology in Eutrope* (AAATE 2009), pages 503-508, August 31 - September 2, Florence (Italy), 2009.

[5] Corpus LDC. Available online: http://www.ldc.upenn.edu/

[6] F. De Rosis, C. Pelachaud, I. Poggi, V. Carofiglio and B. De Carolis, From Greta's mind to her Face: modeling the dynamics of affective states in a conversational embodied agent, *The International Journal of Human-Computer Studies*, 59(1-2):81-118, Elsevier, 2003.

[7] P. Boersma and D. Weenink, Praat: Doing Phonetics by Computer (version 5.0.32), 2008. Available online: http://www.praat.org/

[8] S. Chen, S.A. Billings and W. Luo, Orthogonal least squares methods and their application to non-linear system identification,*International Journal of Control*, 50:1873-1896, 1989.

[9] I. Guyon and A. Elisseeff, An Introduction to variable and feature selection, *Journal of Machine Learning Research*, 3:1157-1182, 2003.

[10] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines and other Kernel-based Learning Methods*, Cambridge University Press, Cambridge, 2000.

[11] Y.C. Ho and R.L. Kashyap, An algorithm for linear inequalities and its applications, *IEEE Trans Electron Comput*, 14(5):683-688, 1965.

[12] The Spider. Available online: http://people.kyb.tuebingen.mpg.de/spider/

[13] I.H Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, 1999. Available online: http://www.cs.waikato.ac.nz/ml/weka