Broadcast News Segmentation with Factor Analysis System

Diego Castan, Alfonso Ortega, Antonio Miguel and Eduardo Lleida

University of Zaragoza, Spain

[dcastan,ortega,amiguel,lleida]@unizar.es

Abstract

This paper studies a novel audio segmentation-by-classification approach based on Factor Analysis (FA) with a channel compensation matrix for each class and scoring the fixed-length segments as the log-likelihood ratio between class/no-class. The system described here is designed to segment and classify the audio files coming from broadcast programs into five different classes: speech (SP), speech with noise (SN), speech with music (SM), music (MU) or others (OT). This task was proposed in the Albayzin 2010 evaluation campaign. The article presents a final system with no special features and no hierarchical structure. Finally, the system is compared with the winning system of the evaluation (the system use specific features with hierarchical structure) achieving a significant error reduction in SP and SN. These classes represent 3/4 of the total amount of the data. Therefore, the FA segmentation system gets a reduction in the average segmentation error rate that is able to be used in a generic task.[1]

Index Terms: Audio Segmentation, Factor Analysis, Broadcast News (BN), Albayzin-2010 Evaluation

1. Introduction

Due to the increase in audio or audiovisual content, it becomes necessary to use automatic tools for different tasks such as analysis, indexation, search and retrieval. Given an audio document, the first step is audio segmentation producing a delineation of a continuous audio stream into acoustically homogeneous regions. When the audio segmentation is followed by a classification system the result is a system that is able to divide an audio file into different predefined classes chosen for a specific task.

Broadcast news (BN) domain is one of the most popular multimedia repositories because it has rich audio types and several approaches have been proposed in this scenario. For example, in the task of automatic transcriptions of BN [1] the data contain clean speech, telephone speech, music segments and speech overlapped with music and noise so the segmentation generates a boundary for every speaker change and environment/channel condition change with no explicit cues. In [2] segmentation is based on five different classes: silence, music, background sound, pure speech, and non-pure speech. The solution is based on SVM combination. In [3] the audio stream from BN domain is segmented into 5 different types including speech, commercials, environmental sound, physical violence and silence. [4] presents a review of different solutions and the acoustic features used in each one of them and also a new algorithm for computing various time-domain and frequency-domain features, for speech and music signals separately, and estimating the optimal speech/music thresholds.

The different segmentation approaches in BN differ in either the feature extraction methods or the classifier. The features can be distinguished in *frame-based* and *segment-based* features. The frame-based features usually describe the signal within a short time period (10-30 ms), where the process is considered stationary. MFCCs or PLPs are commonly used as frame-based features like in [5] where these features are classified with an autoassociative neural network. In [6] the authors propose two pitch-density-based features and relative tonal power density to classify on BN. For segment-based feature extraction, a longer segment is taken into consideration. The length of the segment may be fixed (usually between 0.5 and 5 seconds) or variable. In [7] a content based speech discrimination algorithm is designed to exploit long-term information inherent in modulation spectrum.

Audio segmentation systems perform the segmentation in two different ways. The first one is based on detecting the boundaries and then classifying each delimited segment. We refer to them as *segmentation-and-classification* approaches. For example, in [8], an approach using a temporally weighted fuzzy C-means algorithm has been proposed. The second segmentation way is known as *segmentation-by-classification* and it consists of classifying consecutive fixed-length audio segments. The segmentation is produced directly by the classifier as a sequence of labels. This sequence is usually smoothed to improve the segmentation. An example of this procedure can be seen in [9] where the author combines different features with a GMM and a maximum entropy classifiers. The final sequence-level were smoothed with a HMM.

The different strategies outlined in the preceding paragraphs have their advantages and disadvantages described by Huang and Hansen in [10]. The most common solution to avoid the shortcomings and enjoy the benefits of each strategy is to create hierarchical systems with multiple steps where each level is designed with specific features and segmentation systems for each class. As a result, the system becomes very specific for a database and may produce segmentation errors in different domains. Recently, an audio segmentation task in BN domain was proposed in [11] in the context of the Albayzin-2010 evaluation campaign. Almost all the participants of the evaluation used hierarchical systems, including the winning system [12] based on a hierarchical architecture that used different sets of features for every level.

In this paper, we proposes a whole FA segmentation system with no-hierarchical structure where the within-class variability is compensated with a different channel matrix for each class. The remainder of the paper is organized as follows: database and metric of Albayzin 2010 evaluation is presented in Section 2. Section 3 shows the factor analysis theoretical approach based on FA. Segmentation results are presented in Section 4. Finally, the conclusions are presented in Section 5.

[1] This work has been funded by the Spanish Government and the European Union (FEDER) under the project TIN2011-28169-C05-02

2. Albayzin 2010 audio segmentation evaluation

The Albayzin evaluation campaign is an internationally open set of evaluations organized by the Spanish Network of Speech Technologies (RTTH) every 2 years. A completed description of the Albayzin 2010 evaluation can be found in [13] which describes the participant's approaches and the results of the systems . We summarize the database description and the metric of the evaluation in the next subsections.

2.1. Database

The database consists of a Catalan BN database from the public TV news channel that was recorded by the TALP Research Center from the UPC. It includes approximately 87 hours of annotated audio divided in 24 files of 4 hours long. A set of five different audio classes were defined for the evaluation with the following distribution: Clean speech: 37%; Music: 5%; Speech over music: 15%; Speech over noise: 40%; Others: 3%. The class "Others" is not evaluated in the final test. The database for the evaluation was split into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3).

2.2. Metric

The metric is defined as a relative error averaged over all acoustic classes (ACs):

$$Error = average_i \frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)},$$

where dur($miss_i$) is the total duration of all deletion errors (misses) for the *ith* AC, dur(fa_i) is the total duration of all insertion errors (false alarms) for the *ith* AC, and dur(ref_i) is the total duration of all the *ith* AC instances according to the reference file. The incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. A forgiveness collar of 1 sec (both + and -) is not scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an AC begins/ends.

3. FA-Based audio segmentation

This study proposes a framework for automatic audio *segmentation-by-classification* system. The system deals with the problem of assigning a class label to each fixed-length clips using Factor Analysis (FA) models. The FA approach has been successfully used in speaker recognition [14] [15] [16], speaker verification [17], speaker segmentation [18] and language recognition [19]. The variability of the same class segments is known as *within-class variability*. The goal of these systems is the compensation of the *within-class variability* to reduce the mismatch between training and test. Fig. 1 illustrates the proposed framework where each block is described in the next subsections. We will discuss the feature extraction, the statistic extraction and the within-class variability compensation using FA.

3.1. Acoustic Feature Extraction and Statistics

Mel-frequency cepstral coefficients (MFCCs) [20] are used in most speech recognition tasks because the mel-scale filter bank is an approximation to human auditory system response. Therefore they work well in audio segmention task too. Typically, MFCC features are computed at each short speech segment (e.g., 10 ms) together with their derivatives to capture the short-term speech dynamics. On this framework we extract 16 MFCCs (including C0) computed in 25 ms frame size with a 10 ms frame step, their first and second order derivatives.

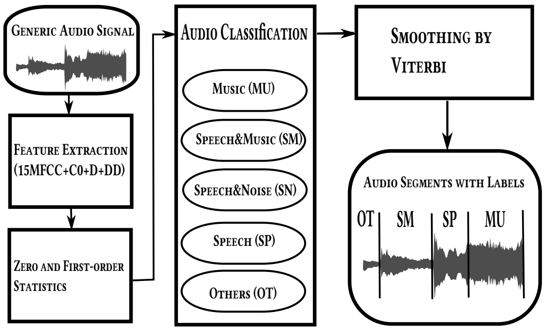

Figure 1: Block Diagram of Factor Analysis Segmentation-by-Classification System for Broadcast News Classes

The audio features are packed in clips of different lengths with 0.1 and 0.5 second clip-steps. The fixed-length clips are mapped to sufficient statistics by using a Universal Background Model (UBM) which is a class-independent GMM with C Gaussians trained with the EM-algorithm [21] on the audio feature vectors of the training data.

3.2. Theoretical Background

Data from a particular class are modeled by a GMM defined by means $m_1, m_2, ..., m_C$, weights $w_1, w_2, ..., w_C$ and covariances $\Sigma_1, \Sigma_2, ..., \Sigma_C$ where C is the number of Gaussians. We can concatenate all GMM means to one mean supervector m of $CF \times 1$ dimensions where F is the feature vector size:

$$m = [m_1^T, m_2^T, ..., m_C^T]^T. \tag{1}$$

The Factor Analysis model is the adaptation of the UBM model where the supervector of means is not fixed and it can vary from segment to segment to account for differences in the channel. These GMMs have segment and class dependent component means but fixed component weights and covariances chosen to be equal to the UBM weights and covariances. Specifically, we use a Factor Analysis model for the mean of kth component of the GMM for segment s:

$$m_k^s = t_k^{c(s)} + U_k x_s \tag{2}$$

where $c(s)$ denotes the class of segment s and $t_k^{c(s)}$ is the channel-independent-class-location vector obtained by using a single iteration of relevance-MAP adaptation from the UBM [22]. U_k is the factor loading matrix and x_s is a vector of L *segment-dependent channel factors* generated by a normal distribution $(N(0, I))$. We stack component-dependent vectors into supervectors m_s and $t^{c(s)}$ and we stack the component-dependent U_k matrices into a single tall matrix U, so that equation can be expressed more compactly as:

$$m^s = t^{c(s)} + Ux_s \tag{3}$$

where U is known as the *channel matrix* and it represents the within-class variability. Note that, following the terminology in the literature, we use the terms *channel matrix* and *channel* *factors* to describe the elements related with the within-class variability even if that variability is not produced by different channels (also can be produced by different speakers or different content). The columns of the U matrix are the basis spanning the subspace of the channel and the *channel factors* are the coordinates defining the position of the channel-dependent supervector in the subspace. The *channel factors* dimension (L) is smaller than CF so U matrix has low rank ($CF \times L$ dimensions). Depending on the application, the value of L is between 50 and 200 and CF can be 98304 if we have 2048 Gaussians and 48-dim feature vector (with the MFCC-UBM). The estimation of these parametres can be understood following [16].

3.3. Class/No-Class U Channel Matrices System

Most of the approaches based on FA for language recognition are implemented with a single U *channel matrix* because the nature of the within-class variability is the same for all the languages as it can be seen in [23] [24] [25] [16]. Therefore, in [26] a segmentation system was proposed with five channel-independent-class-location vectors (one vector per class) and a single compensation channel matrix U for all the classes. The paper compares the FA system with the winnner of the Albayzin-2010 evaluation and the conclusion was that the compensation matrix had a bad behavior for the *Music* class due to the different nature of the rest of the classes. However, the paper [27] shows a clear advantage when the classes are homogeneous (like SN and SP). In this scenario the channel matrix models the compensation between different speakers and different words leaving the background sound as useful information for the classification improving the segmentation.

A number of studies have focused on features to describe the distribution of sounds to be able to distinguish between speech, music or noise. Most of these approaches use a hierarchical structure where each level is specialized on the detection of an specific class with specific features for that class. The main goal of this work is the compensation of all the classes with no-specific features for each class even if the nature of the classes is not the same. We propose a ten channel-independent-class-location vectors (a class and no-class vectors for each class) and five channel matrix representing the within-class variability of each class/no-class with no hierarchical structure. Let

$$\begin{aligned} T =& [t_{mu}, t_{nomu}, t_{ot}, t_{noot}, \\ & t_{sm}, t_{nosm}, t_{sn}, t_{nosn}, t_{sp}, t_{nosp}] \end{aligned} \tag{4}$$

$$\begin{aligned} \Xi \quad = \quad & [U_{mu-nomu}, U_{ot-noot}, \\ & U_{sm-nosm}, U_{sn-nosn}, U_{sp-nosp}] \end{aligned} \tag{5}$$

where T represents the locations of classes and no-classes in the GMM space and Ξ the channel matrices. Our metamodel for class-segment-dependent GMM is parametrized by (T, Ξ) which describe the prior distributions of the parameters m.

This approach will be compared with the classic formulation with a single U *channel matrix* on Section 4 for the classification over the oracle segments and the final segmentation system.

3.4. Scoring

There are different scoring methods used in the state-of-the-art of speaker and language recognition. In the proposed experiments in Section 4 we use the *integration trough the channel factors distributions*. This score is a marginalization using a point estimation of the class m_s, and integrate only over the channel factors, when the statistics are centered around the point estimation m_s. The log-likelihood is defined by the equation (19) in [16] and can be understood following the Section V in the same article.

In [28], [16] and [29] the score employed to detect the speaker is the log-likelihood ratio(LLR). For a test clip χ and class c, the LLR compares the hypothesis that the clip χ belongs to the class c against the hypothesis that the clip χ does not belong to the class c. This score is shown in Formula 6 where the numerator is the likelihood of the test clip calculated with the class model and the denominator is the likelihood of the test clip calculated with UBM model. Note that the UBM model is employed as a general model to describe the not belonging hypothesis. That makes sense for speaker identification task where the hypothesized speaker represents a very small amount into the UBM. However, our problem has only four classes and, therefore, the class is highly represented by the UBM and may corrupt the LLR score.

$$LLR_{class} = log \frac{P(\chi/class)}{P(\chi/UBM)} \tag{6}$$

We propose a LLR scoring where the denominator is the likelihood of the test clip calculated with the no-class model. The compensated log-likelihood ratio (CLLR) is computed for each class/no-class as:

$$CLLR_{class} = log \frac{P(\chi/class)}{P(\chi/noclass)} \tag{7}$$

CLLR is more discriminative than LLR for a segmentation task because the hypothesized class is not presented in the denominator and, also, because the no-class model is channel compensated as the class model.

4. Experimental results

In a segmentation-by-classification system, the errors can be produced in two ways: first, a classification error due to a bad labeled frame, and a segmentation error due to a temporal mismatch between the oracle boundaries and the hypothesis boundaries. This Section shows the experiments for the evaluation data described in Section 2.1 divided into two sets. In the first set, the segments are given by the ground truth and the systems decide the class of each segment with no segmentation error to evaluate the classification accuracy of the systems. The second set of experiments shows the segmentation and the classification error and it proposes a final segmentation-by-classification system based on FA that improve the result of the winning system in the Albayzin evaluation.

4.1. Classification Experiments with Oracle Segmentation

The classification is done over the segments extracted with the ground truth to evaluate the classification accuracy over the whole segment. Most of the segments are between 5 and 20 seconds long.

We propose two sets of systems based on GMM and HMM-GMM as a baseline. Table 1 shows the results for these systems. In the first part of the table, we have tried with different number of Gaussians. The classification is based on the highest accumulated likelihood over the whole segment. Increasing the number of Gaussians improves the final result. The highest

Table 1: Classification Baseline Experiments: error per class and total error for GMM-HMM systems over the test files with perfect segmentation in %

GMM	MU	SP	SM	SN	**TOTAL**
32G	9.66	49.36	37.59	48.11	**36.18**
64G	10.68	45.74	36.68	45.44	**34.63**
128G	9.81	41.79	32.02	40.75	**31.09**
256G	10.43	37.61	31.85	37.67	**29.39**
512G	9.51	35.95	29.38	35.99	**27.71**
1024G	9.39	34.91	27.03	34.35	**26.42**
2048G	9.61	33.39	38.01	34.01	**26.25**
HMM-LeftToRight	MU	SP	SM	SN	**TOTAL**
1 ST - 2048G	9.61	33.39	28.01	34.01	**26.25**
2 ST - 1024G	9.48	42.75	27.45	41.26	**30.24**
4 ST - 512G	10.11	27.91	27.17	29.87	**23.77**
8 ST - 256G	8.37	31.64	26.42	32.1	**24.63**
16 ST - 128G	8.84	26.92	32.28	32.12	**25.04**
32 ST - 64G	11.33	29.81	26.64	32.48	**25.07**

number of Gaussians is 2048 because, although the final results is the best one, the MU and SM classes begin to get worse results. The next experiment of the baseline system uses 2048 Gaussians distributed in different nodes in a HMM. The second part of the Table 1 shows the results for left-to-right topologies of HMMs. These topologies increase the activity duration of each model [30], avoiding wrong transitions inside the segment and improving the results. The best baseline system (23.77% of total error) is performed using five left-to-right HMMs with four emitting states and 512 Gaussians per state where each HMM corresponds to one acoustic class.

To evaluate the strengths and weaknesses of a FA system, we assess different configurations described in Section 3. The UBM employed to compute the statistics has a fixed amount of 2048 Gaussians to be able to compare the results of the FA systems with the GMM/HMM baseline. We compute the result over the test set using the *integration trough the channel factors distributions scoring*. The experiments are calculated with a single channel matrix to compensate all the classes and different channel matrices for each class/no-class using different number of channel factors (50, 100, 150, 200 and 250).

Table 2: FA systems with a single U for all the classes and U matrix for every class/no-class over the test set with perfect segmentation in %

Single U	MU	SP	SM	SN	**TOTAL**
50 chnf	10.20	15.98	24.21	21.41	**17.95**
100 chnf	9.16	16.06	20.28	20.06	**16.39**
150 chnf	9.42	15.52	18.04	18.90	**15.47**
200 chnf	9.08	15.72	17.38	19.17	**15.34**
250 chnf	8.52	16.70	16.06	19.42	**15.17**
U per class	MU	SP	SM	SN	**TOTAL**
50 chnf	9.65	19.13	24.10	23.31	**19.05**
100 chnf	8.54	16.22	22.12	20.18	**16.77**
150 chnf	9.65	16.63	18.31	19.49	**16.02**
200 chnf	9.20	17.22	17.73	19.60	**15.94**
250 chnf	9.69	17.46	17.12	19.82	**16.02**

Comparing the Table 1 and the Table 2, it can be seen a significant improvement using FA as a classification system against GMM/HMMs. Using the best HMM configuration (left-to-right HMM with four states and 256 Gaussians in each state) as a reference, the worst FA system improves the total result in 4.72% (with a U matrix per class and 50 channel factors) and in 8.6% comparing with the best FA configuration (with a single U matrix and 250 channel factors).

4.2. Segmentation-by-Classification Experiments

In the last subsection, each segment was labeled with the best decision coming from the accumulated log likelihood or accumulated log likelihood ratio of the models. In this subsection, the segments are delimited with the transitions between the scores and the errors might be due to a temporal mismatch or a bad label assignment.

Table 3: Segmentation Baseline Experiments: error per class and total error for HMM systems over the test files in %

HMM-LeftToRight	**MU**	**SP**	**SM**	**SN**	**TOTAL**
1 ST - 2048G	35.53	59.22	65.07	58.60	**54.6**
2 ST -1024	29.96	59.26	54.79	56.82	**50.21**
4 ST - 512	26.04	49.8	45.98	50.27	**43.02**
8 ST - 256G	24.35	49.3	41.66	50.19	**41.37**
16 ST -128G	17.82	40.24	36.02	43.06	**34.28**
32 ST - 64G	17.39	39.53	33.95	41.56	**33.31**

As we did in the last subsection, GMM/HMM systems are used as the baseline. Because the segments are delimited by the scoring transitions, the scores need to be smooth using low pass filters or HMM. Table 3 shows different HMM topologies and configurations. Again, the left-to-right topology improves the result because these systems smooth the transitions between classes. The best baseline system for segmentation-by-classification (33.31% of total error) has 32 states with 64 Gaussians each state and has a left-to-right topology.

Table 4: FA segmentation-by-classification systems with a single U for all the classes and U matrix for every class/no-class in %

	Win-3.0 step-0.5 100chnf				
	MU	SP	SM	SN	**TOTAL**
Single U	40.38	76.91	60.52	64.31	**60.53**
U per class	33.35	45.62	36.2	47.44	**40.65**

As a preliminary experiment, the first FA segmentation-by-classification system computes the statistics over a 3 second windows with 0.5 second window-steps and 100 channel factors. An increment of the channel factors or a reduction of the window-step increase the memory and the time to train the models exponentially. Experiments with a single channel matrix for all the classes and a channel matrix for each class are presented in Table 4. There is a significant improvement in the majority classes using a channel matrix for each class because the CLLR removes the information of the target class in the denominator as we pointed in Section 3.4. The bigger is the class in the data, more significant is the reduction of the error comparing with a single channel matrix for all the classes. Accordingly, the total error is reduced about 20%.

Once determined that the best configuration is the FA system with a channel matrix for each class, the window-step can be modified to get more resolution (0.1 second window-step) and the CLLR can be smoothed to avoid an over segmentation. In the experiments, a zero-phase average filter is computed to smooth the CLLR of each class and avoid a sudden change in the segment labels. Figure 2 shows the filtered-ratio scores for each class over a chunk of a test file. The ground truth is plotted in the same figure and it is represented with a square wave of

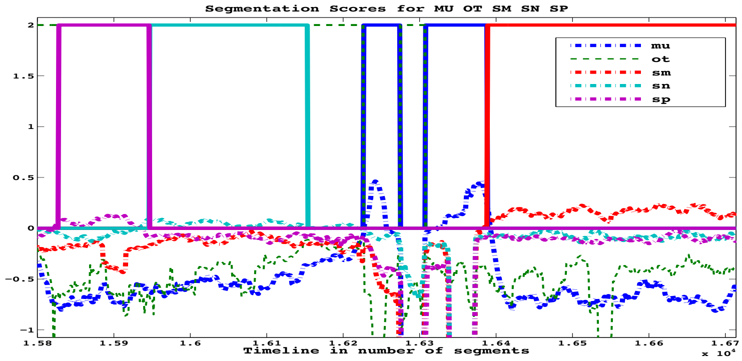

Figure 2: Scores and the ground truth of each class over a chunk of a test file

amplitude 2. The color of each score class and the corresponding ground truth is the same. The figure clearly shows that the ratio of the winning class is bigger than zero and corresponds with the ground truth class.

Due to the metric, the smallest classes have to be detected with the same accuracy as the largest classes as can be seen in section 2.2. To increase the detection of the smallest class (MU) we optimize the prior probabilities in a Viterbi algorithm checking the total result over the train files. Table 5 shows the total error over the train files. The first row shows the total error when all the classes have the same priority and it can be seen that the smallest error is obtained when MU and SN/SP have 28% and 16% of priority respectively decreasing the false alarms of the SN/SP over MU class. These priors are employed in the Viterbi over the test files and the results are shown in Table 6.

We compare the error of the system proposed in this work with the winning system of the Albayzin-2010 evaluation [12] where 15 MFCCs, the frame energy, and the derivatives are extracted. In addition, the spectral entropy and the Chroma coefficients are calculated. The mean and variance of these features are computed over 1 second interval creating a 122 dimension feature vectors. The segmentation approach chosen is HMM-based. The acoustic modeling is performed using five HMMs with three emitting states and 256 Gaussians per state. Each HMM corresponds to one acoustic class. A hierarchical organization of binary HMM detectors is used. First, audio is segmented into Music/non-Music portions. Second, the non-Music portions are further segmented into Speech-over-music/non-Speech-over-music portions. Finally, the non-Speech-over-music portions are segmented into Speech/Speech over noise.

Table 6 is divided in two parts: the first part shows the error for each class and the average error for the winning herarchical-HMM system of the evaluation (HMM-Winn). The last column shows the NIST metric used in the NIST RT Diarization evaluations [31] to compare the systems with a well-known metric. To be able to compute the NIST error with the herarchical-HMM system, we replicate the winning system according to [12] (HMM-Rep). The second part of the table shows FA segmentation-by-classification system (FA-Segm) after the Viterbi smoothing with the priors of the Table 5. The last row of the table shows the same FA system with a slight modification introducing OT segments between SN and SP to model the silence of the anchor before the coverage to avoid the false alarms. The hierarchical-HMM systems detects better the MU and SM segments than the FA systems due to the Chroma coefficients in the features. However, SN and SP classes are much better detected with the FA system decreasing the error of the classes in 2% and 9% respectively. These classes represent more that 3/4 of the total amount of the data, therefore the classification of the total time is also increased substantially. The FA systems reduces the average error in a 2% with the Albayzin metric and almost 3% with the NIST metric.

Table 5: Results over the train files to select the priors for each class in %

Prior of each class					AVG Error
MU	OT	SM	SN	SP	over the train files
0.20	0.20	0.20	0.20	0.20	15.95%
0.22	0.20	0.20	0.19	0.19	14.52%
0.24	0.20	0.20	0.18	0.18	13.75%
0.26	0.20	0.20	0.17	0.17	13.39%
0.28	0.20	0.20	0.16	0.16	**13.23%**
0.30	0.20	0.20	0.15	0.15	13.25%

Table 6: Error per class and total error for Albayzin evaluation winning system and Factor Analysis Segmentation system over the test files in %

	Error for each class					
	MU	SM	SN	SP	TOTAL	NIST
HMM-Winn	19.2	25.0	37.2	39.5	30.2	-
HMM-Rep	16.3	24.0	38.8	40.8	30.0	19.3
FA-Segm	21.7	27.6	35.4	30.5	28.8	**16.9**
FA-Segm OT	21.7	27.6	34.0	29.5	**28.2**	17.5

5. Conclusion

This paper describes a new segmentation-by-classification system based on Factor Analysis approach. The system has been applied for the segmentation of BN. The task consists of the segmentation of audio files and further classification into 5 different classes as proposed in the Albayzin 2010 evaluation. The solution we propose here compensates the within-class variability creating a channel matrix for each class and scoring the segments as the ratio between class/no-class. This approach has been compared with HMM-GMM baseline systems and with the winning system of the evaluation showing a significant improvement in both cases even if the best results in the evaluation were obtained by an HMM/GMM based hierarchical system that made use of MFCC along with Chroma features. Experimental results show that the FA approach allows a significant reduction in the classification of SP and SN and thus a reduction in the average segmentation error rate.

6. References

[1] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Workshop*, 1998.

[2] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, Apr. 2003.

[3] T. Nwe and H. Li, "Broadcast news segmentation by audio type analysis," *Acoustics, Speech, and Signal Processing, 2005 . . .*, vol. 2, pp. ii–1065, 2005.

[4] Y. Lavner and D. Ruinskiy, "A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, pp. 1–15, 2009.

[5] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of audio signals using AANN and GMM," *Applied Soft Computing*, vol. 11, no. 1, pp. 716–723, Jan. 2011.

[6] L. Xie, Z.-H. Fu, W. Feng, and Y. Luo, "Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news," *Multimedia Systems*, vol. 17, no. 2, pp. 101–112, Sep. 2010.

[7] M. Markaki and Y. Stylianou, "Discrimination of speech from nonspeeech in broadcast news based on modulation frequency features," *Speech Communication*, vol. 53, no. 5, pp. 726–735, May 2011.

[8] N. Nguyen, M. Haque, C.-h. Kim, and J. Kim, "Audio segmentation and classification using a temporally weighted fuzzy C-means algorithm," *Advances in Neural Networks . . .*, pp. 447–456, 2011.

[9] A. Misra, "Speech/Nonspeech Segmentation in Web Videos," *research.google.com*, 2012.

[10] R. Huang and J. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," *Audio, Speech, and Language . . .*, vol. 14, no. 3, pp. 907–919, 2006.

[11] T. Butko, C. N. Camprubí, and H. Schulz, "Albayzin-2010 audio segmentation evaluation: evaluation setup and results," in *FALA Evaluation*, 2010, pp. 305–308.

[12] A. G. Antolín and R. S. S. Hernández, "UPM-UC3M system for music and speech segmentation," in *Proc. II Iberian SLTech*, 2010, pp. 421–424.

[13] T. Butko and C. Nadeu, "Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, p. 1, 2011.

[14] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, May 2005.

[15] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," *Online: http://www. crim. ca/perso/patrick. kenny*, pp. 1–17, 2006.

[16] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[17] C. Vaquero, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification," in *Proc Interspeech 2010*, vol. 2010, 2010, pp. 2310–2313.

[18] C. Vaquero, A. Ortega, and E. Lleida, "Intra-session variability compensation and a hypothesis generation and selection strategy for speaker segmentation," *Acoustics, Speech and Signal . . .*, pp. 3–6, 2011.

[19] N. Brummer, A. Strasheim, V. Hubeika, P. Mat\vejka, L. Burget, and O. Glembek, "Discriminative acoustic language recognition via channel-compensated GMM statistics," in *Tenth Annual Conference of the International Speech Communication Association*, 2009, pp. 2187–2190.

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal . . .*, no. 4, 1980.

[21] C. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006, vol. 4.

[22] D. a. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[23] H. Li, B. Ma, and K. Lee, "Spoken Language Recognition: from Fundamentals to Practice," *Proceedings of IEEE*, 2013.

[24] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1969–1978, Sep. 2007.

[25] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, Jan. 2008.

[26] D. Castan, A. Ortega, and E. Lleida, "Factor Analysis Segmentation and Classification in Broadcast News Domain," in *Proc. III Iberian SLTech*, 2012.

[27] D. Castan, C. Vaquero, A. Ortega, D. Martínez, and E. Lleida, "Hierarchical Audio Segmentation with HMM and Factor Analysis in Broadcast News Domain," in *Interspeech*, 2011.

[28] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with Joint Factor Analysis," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Ieee, Apr. 2009, pp. 4057–4060.

[29] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP*, vol. 1. Citeseer, 2005, pp. 637–640.

[30] J. Bilmes, "What HMMs can do," *Graphical Models*, no. 206, 2002.

[31] NIST, "The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan," pp. 1–18, 2009.