# SEGMENTATION-BY-CLASSIFICATION SYSTEM BASED ON FACTOR ANALYSIS

*Diego Castan, Alfonso Ortega, Jesus Villalba, Antonio Miguel, Eduardo Lleida* *

ViVoLab I3A
University of Zaragoza, Spain
dcastan,ortega,villalba,amiguel,lleida@unizar.es

**ABSTRACT**

This paper proposes a novel audio segmentation-by-classification system 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 scores are smoothed and the most probable sequence is computed with a Viterbi algorithm. 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 system is compared with the winning system of the evaluation achieving lower error rate 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.

***Index Terms***— Audio Segmentation, Factor Analysis, Channel Compensation, 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. In [5], a system of three components (segmentation, clustering and classification) is used to recognize an entire half an hour show with no prior knowledge of acoustic conditions and speakers.

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 [6] where these features are classified with an autoassociative neural network. In [7] 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 [8] 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 [9], 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 [10] where the author combines different features with a GMM and a maximum entropy classifiers. The final sequence-level were smoothed with a HMM.

An audio segmentation task was proposed [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. For this evaluation database, in [13] we proposed a system that uses a 2-level hierarchical architecture where the second level is based on FA minimizing the segmentation error over this database.

In this paper, we proposes a whole FA segmentation system 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.

*This work has been funded by the national 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. Nowadays, the quantitative comparison and evaluation of competing approaches is very important in nearly every research and engineering problem. The evaluation campaigns that independently compare systems from different research groups help us to determine which directions are promising and which are not. A completed description of the Albayzin 2010 evaluation can be found in [14] 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.

Five different audio classes were defined for the evaluation: music (MU), speech (SP), speech with music (SM), speech with noise (SN) and others (OT) but this class is not evaluated in final test. The distribution of the classes within the database is the following: Clean speech: 37%; Music: 5%; Speech over music: 15%; Speech over noise: 40%; Other: 3%.

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.

The proposed metric is slightly different from the conventional NIST metric for speaker diarization, where only the total error time is taken into account independently of the acoustic class. Since the distribution of the classes in the database is not uniform, the errors from different classes are weighed differently (depending on the total duration of the class in the database). Therefore the participants have to detect correctly not only the best-represented classes (speech and speech over noise, 77% of total duration), but also the minor classes (like music, 5%).

## 3. SEGMENTATION SYSTEM BASED ON FACTOR ANALYSIS

The Factor Analysis approach has been successfully used in speaker recognition [15] and more recently in language recognition [16]. In these tasks, the main problem is the session variability given by different channel conditions in training and testing. The goal of these systems is the compensation of the channel variability between different utterances due to speakers, background noises or recording devices presented in the audio.

This work deals with the problem of assigning a class label to each fixed-length segment using FA models trying to compensate the within-class variability. At the first stage, we extract 16 MFCCs (including C0) computed in 25 ms frame size with a 10 ms frame step, their $\Delta$ and $\Delta\Delta$ for every file of the database. The audio features are packed in 3 second segments with 0.1 second segment step and each segment is described as statistics over a Universal Background Model (UBM). The statistic extraction is described more precisely in the next subsection. The classifier with the channel compensation and the scoring method proposed in this work are described in subsections 3.2 and 3.3 respectively. The segmentation is produced directly by the classifier as a sequence of decisions. Additionally, a smoothing with an average filter and a Viterbi algorithm is required to find in a recursive manner the most probable sequence under the assumption that a sudden change of sound types in an arbitrary way is unlikely. The parameters of the filter and the priors for the Viterbi algorithm are described in section 4.

### 3.1. Statistics

The fixed-length segments are mapped to sufficient statistics by using a Universal Background Model (UBM) which is a class-independent GMM with 2048 Gaussians trained with the EM-algorithm on the audio feature vectors of the training data. Following the classic terminology of the bibliography, we refer mean-vector and diagonal precision matrix of the UBM as $\mu_k$ and $\Sigma_k$ where $k$ is the Gaussian component index. All further processing is based only on the statistics, rather than the original feature vectors. Let $P_{ksi} = P(k|\phi_{si})$ denote the posterior probability of UBM component $k$, given feature vector $\phi_{si}$, computed with the standard method for GMM observations, assuming frame-independence. For segment s, with feature vectors indexed $i = 1, 2, ..., N_s$, we define the zero and first-order statistics respectively as:

$$n_{sk} = \sum_{i=1}^{N_s} P_{ksi}$$

$$f_{sk} = \sum_{i=1}^{N_s} P_{ksi} \Sigma_k^{1/2} (\phi_{si} - \mu_k)$$

We center and reduce our statistics relative to the UBM. After this transformation the formulas below can be written with the statistic terms.

### 3.2. Theoretical Background

Data from a particular class segment 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. 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:

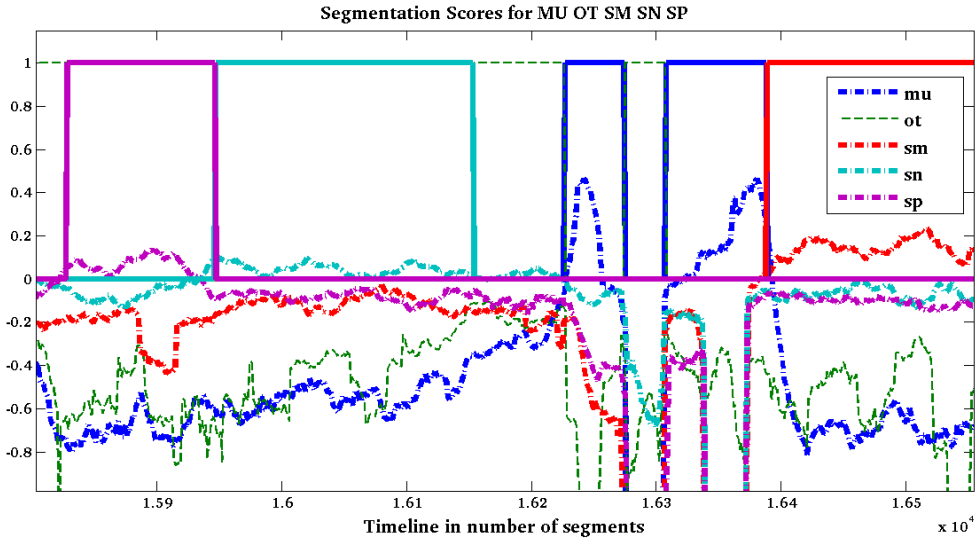


**Fig. 1**. Scores and the ground truth of each class over a chunk of a test file

$$m_{sk} = t_{c(s)k} + U_k x_s$$

where $c(s)$ denotes the class of segment $s$ and $t_{sk}$ is the channel-independent-class-location vector obtained by using a single iteration of relevance-MAP adaptation from the UBM. This adaptation is expressed in term of statistics as:

$$t_{ck} = \frac{\sum_s f_{sk}}{r + \sum_s n_{sk}}$$

where the sums are over all segments $s$ belonging to the class $c$ and $r$ is the relevance factor ($r = 14$ in our experiments). $U_k$ is the factor loading matrix which is the subspace of channel variability and $x_s$ is a vector of L segment-dependent channel factors generated by a normal distribution. Channel factor vector $x_s$ can be seen as the coordinates of the channel dependent class segment vector in the subspace defined by $U_k$ . 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:

$$m_s = t_{c(s)} + U x_s$$

where $U$ is known as the channel matrix and it represents the within-class variability.

The parameter $U$ can be estimated using the EM algorithm iteratively. Data from many segments are used, where the channel factors of each segment is treated as a hidden variable. In the E-step posterior distributions of $x$ are estimated for each segment, using current parameters as:

$$\hat{x}_s = \left(I + \sum_k n_{sk} U_k' U_k\right)^{-1} U_k' f_s.$$

In the M-step we find parameters $U$ that maximize an auxiliary function involving the old and the new parameters. Understanding the training process of U channel matrix can be complex so we defer the responsibility of this algorithm following [15].

### 3.3. System description

In [17] a 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 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. In the speech classes (SM, SN and SP), the channel matrix is modeling the compensation between different speakers and different words leaving the background sound as useful information for the classification. This model is totally different for the music class (MU) because the channel matrix should model the compensation between different tones and instruments.

The main goal of this work is the compensation of all the classes 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. Let

$$T = [t_{mu}, t_{nomu}, t_{ot}, t_{noot}, t_{sm}, t_{nosm}, t_{sn}, t_{nosn}, t_{sp}, t_{nosp}]$$

$$\Xi = [U_{mu-nomu}, U_{ot-noot}, U_{sm-nosm}, U_{sn-nosn}, U_{sp-nosp}]$$

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

In [18], different scoring methods are studied. In this approach, the *integration trough the channel factors distributions* is done. We compute the log-likelihood of each segment with respect to every class/no-class. Finally, the detection log-likelihood ratio is computed for each class/no-class as:

$$Ratio_{detclass} = logP(s/class) - logP(s/noclass)$$

A zero-phase average filter is computed to smooth the ratio of each class and avoid a sudden change in the segment labels. Different filter lengths have been computed over the scores but we concluded that a two samples average filter is adequate to smooth the

decisions. Figure 1 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 amplitude 1. 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.

## 4. EXPERIMENTAL RESULTS

| | **Error for each class** | | | | |
|---|---|---|---|---|---|
| | MU | SM | SN | SP | **AVG** |
| FA-Class. | 29.3% | 29.1% | 29.6% | 22.4% | **27.6%** |

**Table 1**. Error per class and total error for Factor Analysis system over the test files with oracle boundaries

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. To show the accumulated error of the system, Table 1 shows the error of each class when classification is produced for oracle segmentation. Each oracle segment is labeled as one class. The winning class is the one that presents the highest accumulated score along each considered segment.

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 the Viterbi algorithm checking the total result over the train files. Table 2 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 15% 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 3.

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 $\Delta$ and $\Delta\Delta$ 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. 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.

| **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 2**. Results over the train files to select the priors for each class

Table 3 shows the error for each class and the average error for the Albayzin evaluation winning system and the Factor Analysis Segmentation system proposed in section 3. The HMM-Hierarchical system detects better the MU and SM segments than the FA system 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. Using the metric proposed in the evaluation, the FA system reduces the average error in more than 1%. Also, comparing the classification oracle segment error in table 1 with the total segmentation error in table 3, it can be concluded that the delimitation of the boundaries is quite acceptable because the total error has increased only 1.5%.

| | **Error for each class** | | | | |
|---|---|---|---|---|---|
| | MU | SM | SN | SP | **AVG** |
| HMM-Hierar. | 19.2% | 25.0% | 37.2% | 39.5% | 30.2% |
| FA-Segm. | 22.8% | 27.6% | 35.4% | 30.5% | **29.1%** |

**Table 3**. Error per class and total error for Albayzin evaluation winning system and Factor Analysis Segmentation system over the test files

## 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 that took place in the conference FALA 2010 organized by the Spanish Network on Speech Technologies. The best results in the evaluation were obtained by an HMM/GMM based hierarchical system that made use of MFCC along with Chroma features. 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. 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] Chen SS. and Gopalakrishnan PS., "Ibm l csr system for transcription of broadcast news used in the 1997 hub4 english evaluation," in *Proceedings of the Speech Recognition Workshop, 1998*, 1998.

[2] Lu L, Zhang H-J, and Li SZ., "Content-based audio classification and segmentation by using support vector machines," in *Multimedia Systems*, 2003.

[3] Nwe TL and Li H., "Broadcast news segmentation by audio type analysis," in *IEEE International Conference in Acoustics, Speech, and Signal Processing*, 2005.

[4] Y Lavner and A Ruinskiy, D, "Decision-tree-based algorithm for speech/music classification and segmentation," in *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[5] Siegler MA, Jain U., Raj B., and Stern RM., "Automatic segmentation, classification and clustering of broadcast news audio," in *Signal Processing*, 2009.

[6] Palanivel S. Dhanalakshmi, P. and V. Ramalingam, "Classification of audio signals using aann and gmm," in *Applied Soft Computing*, 2011.

[7] L. Xie, Z. Fu, and Y. Feng, W.and Luo, "Pitch-density-based features and an svm binary tree approach for multi-class audio classification in broadcast news," in *Multimedia Systems*, 2010.

[8] M. Markaki and Y. Stylianou, "Discrimination of speech from nonspeeech in broadcast news based on modulation frequency features," in *Speech Communication*, 2011.

[9] Haque M. Kim C. Nguyen, N. and J. Kim, "Audio segmentation and classification using a temporally weighted fuzzy c-means algorithm," in *Advances in Neural Networks*, 2011.

[10] A. Misra, "Speech/nonspeech segmentation in web videos," in *Research Google*, 2012.

[11] Butko T, Nadeu C, and Schulz H., "Albayzin-2010 audio segmentation evaluation: Evaluation setup and results," in *FALA Segmentation Evaluation*, 2010.

[12] Gallardo A and San Segundo R., "Upm-uc3m system for music and speech segmentation," in *Proc. FALA.*, 2010.

[13] Castan D., Vaquero C, Ortega A, Martinez D, and Lleida E., "Hierarchical audio segmentation with hmm and factor analysis in broadcast news domain," in *Interspeech*, 2011.

[14] T. Butko and C. Nadeu, "Audio segmentation of broadcast news in the albayzin-2010 evaluation: overview, results, and discussion," in *EURASIP Journal on Audio, Speech, and Music Processing*, 2011.

[15] Kenny P, Boulianne G, Ouellet P, and Dumouchel P., "Joint factor analysis versus eigenchannels in speaker recognition," in *IEEE Transactions on Audio, Speech and Language Processing*, 2007.

[16] Brummer N, Strasheim A, and et al. Hubeika V, "Discriminative acoustic language recognition via channel-compensated gmm statistics," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[17] Castan D., Ortega A., and Lleida E., "Factor analysis segmentation and classification in broadcast news domain," in *Proceedings IberSpeech*, 2012.

[18] Glembek O, Burget L, Dehak N, Brummer N, and Kenny P., "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.