Hierarchical Audio Segmentation with HMM and Factor Analysis in Broadcast News Domain

Diego Castán, Carlos Vaquero, Alfonso Ortega, David Martínez, Jesús Villalba, and Eduardo Lleida

Communication Technology Group (GTC)
Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain
[dcastan,cvaquero,ortega,david,jvillalba,lleida]@unizar.es

Abstract

This paper investigates the performance of a Factor Analysis stage in audio segmentation systems. The system described here is designed to segment and classify the audio files coming from broadcast programs into five different classes: speech, speech with noise, speech with music, music or others. This task was recently proposed as a competitive evaluation organized by the Spanish Network on Speech Technologies as part of the conference FALA 2010. The system proposed here makes use of a hierarchical structure in two steps with two different acoustic features. First, the system decides among music, speech with music or the rest of the classes by using HMM/GMM and a smoothed combination of MFCC and Chroma as feature vectors. Next, the system classifies speech and speech with noise by using FA and MFCC as acoustic features. The results shows that, with this configuration, the error rate achieved is lower than the one obtained by the best system presented in the FALA 2010 evaluation.

Index Terms: Factor analysis, Intersession variability compensation, Broadcast segmentation, Chroma features, Hidden Markov Models

1. Introduction

Audio segmentation is the task of delineating a continuous audio stream in terms of acoustically homogeneous regions. Segmentation plays an important role in audio processing applications and it has received increasing attention for its applications in content-based audio retrieval recognition and classification. An audio segmentation and classification system divides an audio file into different segments where each segment or clip should consist of a single class that is acoustically different from other classes of the audio file. A good segmentation should be able to delimit the boundaries between two classes to group segments into homogeneous classes. This is very useful for many other systems. For example, a previous identification of speech segments facilitates the task of speech recognition or speaker diarization. In addition, audio segmentation is widely used to make online adaptation of ASR models.

Different research groups work on audio segmentation in many scenarios. In [1] segmentation is based on five different classes: silence, music, background sound, pure speech, and non-pure speech which include speech with music and speech with noise. The proposed solution is based on a combination of SVM. In [2] the audio stream from broadcast news domain is segmented into 5 different types including speech, commercials, environmental sound, physical violence and silence. One of the most important tasks in audio segmentation is the speech-music segmentation. A review of different solutions and the acoustic features used in each one them can be found in [3]. Also it presents a new algorithm which consists of a learning phase with predefined training data used for computing various time-domain and frequency-domain features. The goal is to segregate speech from music estimating the optimal speech/music thresholds, based on the probability density functions of the features. An automatic procedure is employed to select the best features for separation based on a tree algorithm. It can be found that many of the proposed solutions use the same features by varying the length of the feature vectors and the classification algorithm.

In the context of the Albayzin-2010 evaluation campaign, which is an internationally-open set of evaluations organized by the Spanish Network on Speech Technologies, an audio segmentation task was proposed in [4]. For this task, we propose a system that uses a 2-level hierarchical architecture: the first level is a segmentation system based on HMM-GMM to classify classes that contains music. The second level is the novel aspect of our proposal. This level is based on Factor Analysis and performs a classification over the segments delimited by a Bayesian Information Criterion (BIC). Also, a Gaussian Back End is applied over the scores of the Factor Analysis models. The main reason that allows the improvement of the best results of the Albayzin-2010 evaluation campaign, is the compensation of different speakers and channels in the classification of speech classes, denoted as inter-session compensation following the terminology of speaker recognition. Each level uses a specific feature set based on the target classes.

The remainder of this paper is organized as follows: database and metric of Albayzin 2010 audio segmentation evaluation is presented in Section 2, Section 3 discusses the set of features and the system algorithm description, Section 4 provides an evaluation of the systems and is followed by the conclusion in Section 5.

2. Albayzin 2010 audio segmentation evaluation

2.1. Database

The database used for Albayzin-2010 segmentation evaluations consists of a Catalan broadcast news database from the public TV news channel that was recorded by the TALP Research Center from the UPC, and was manually annotated by Verbio Technologies. The database includes around 87 hours of annotated audio (24 files of approximately 4 hours long).

Five different audio classes were defined for the evaluation:

1. Music (MU): Music is understood in a general sense. Most of the segments belonging to this class are part of

the opening music of the news programs and music from different reports.

2. Speech (SP): Clean speech in studio from a close microphone.
3. Speech with music in background (SM): Overlapping of speech and music classes or speech with noise in background and music classes.
4. Speech with noise in background (SN): Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation).
5. Other (OT): This class refers to any type of audio signal (including noises) that does not correspond to the other four classes. 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 evaluation was splitted into 2 parts: for training (2/3 of the total amount of data), and testing (the remaining 1/3). The audio signals are provided in PCM format, mono, 16 bit resolution, and sampling frequency 16 kHz.

2.2. Metric

The metric is defined as a relative error averaged over all acoustic classes (ACs) as proposed in [4] for the Albayzin 2010 evaluation:

$$Error = average_i \frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}, \qquad (1)$$

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 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 weighted differently (depending on the total duration of the class in the database). Therefore the participants have to detect accurately not only the best-represented classes (speech and speech over noise, 77% of total duration), but also the minor classes (like music, 5%).

3. System description

3.1. Acoustic Features

This section is a summary of the acoustic feature extraction method used in this system. The inputs to train the segmentation system are segments of varying duration. For our purpose, we use two different acoustic features: the first one is developed to detect music classes with an HMM/GMM approach and the second one is intended to differentiate between speech classes using a FA strategy.

For the first acoustic feature vector type, the one used by the HMM/GMM steps, we extract 16 MFCC (including C0) computed in 25ms frame size with a 10ms frame step, their delta and double delta. Also, a 12 chroma feature vector is concatenated to improve the detection of music frames as it is described in [5]. Chroma features are extracted on 64ms frame size with a 10ms frame step. After that, the mean and standard deviation are computed over 1 second windows with an overlap of 0.5 seconds. Thus, the system uses 120 features (60 for the mean and 60 for the standard deviation of the MFCC and the Chroma features) every 0.5 second.

The second feature vector, the one used by the FA stage, is composed of 16 MFCC (including C0) computed in 25ms frame size with a 10 ms frame step, their delta and double delta. A representation of our feature extractor is shown in Fig. 1.

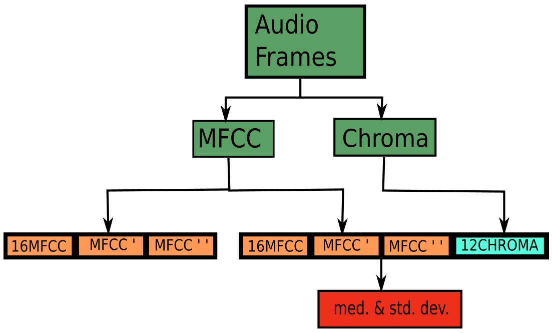

Figure 1: *Feature extraction*

3.2. Segmentation and Classification

This section describes our system as a 2-level hierarchical architecture. For the first level we have considered a 7-state HMM for each acoustic class. The number of states has been adjusted from preliminary experiments. In [5] a more detailed study can be found where different aspects of an HMM/GMM classification system developed for the same task and database are discussed. The second level of our hierarchical system is based on factor analysis used recently in language recognition [6] and speaker recognition [7]. A 1024-component maximum-likelihood GMM was trained with the EM-algorithm on the acoustic feature vectors of all available classes to build the Universal Background Model, or UBM. We shall refer respectively to the mean-vector and (diagonal) precision matrix of Gaussian component k of the UBM as μ_k and Λ_k. All input sequences, for both training and test purposes, are mapped to sufficient statistics and all further processing is based only on the statistics, rather than the original feature sequences. Let $P_{ksi} = P(k|\phi_{si})$ denote the posterior probability of UBM component k, given the feature vector ϕ_{si}, computed with the standard recipe for GMM observations, assuming frame-independence. For segment s, with frames indexed $i = 1, 2, ..., N_s$, we define the zero and first-order statistics respectively as:

$$n_{sk} = \sum_{i=1}^{N_s} P_{ksi} \qquad (2)$$

$$f_{sk} = \sum_{i=1}^{N_s} P_{ksi} \Lambda_k^{1/2} (\phi_{si} - \mu_k), \quad (3)$$

where $k = 1, ..., 1024$. For later convenience, we stack the first-order vectors for all components into a single supervector, denoted as f_s. We also center and reduce our statistics relative to the UBM, so that we can henceforth regard the UBM as having zero mean and unity precision for all components. This simplifies the formulas below for working with the statistics, because after this transformation we do not need to refer to the UBM parameters again.

The Factor Analysis model has a two-level hierarchy: first, we assume there is a different Gaussian mixture model (GMM) that generates every observed speech segment. Second, we assume a metamodel that generates the GMM for every segment. These GMMs have segment and class dependent component means, but fixed component weights and precisions, chosen to be equal to the UBM weights and precisions. Specifically, we use a Factor Analysis model for the kth component mean of the GMM for segment s:

$$m_{sk} = t_{c(s)k} + U_k x_s, \quad (4)$$

where $c(s)$ denotes the class of segment s; the t_{sk} are the class location vectors; x_s is a vector of C segment-dependent channel factors; and U_k is the factor loading matrix. As in the case of the first-order statistics, we stack component-dependent vectors into supervectors m_s and t_c and we stack the component-dependent U_k matrices into a single tall matrix U, so that 3 can be expressed more compactly as:

$$m_s = t_{c(s)} + U x_s \quad (5)$$

where U is known as channel matrix and it represents the within-class variability. For our system, we use a 100 channel factors. Let $T = [t_{sn} t_{sp}]$ where T represents the locations of classes in GMM space, so our metamodel for class-segment-dependent GMMs is parameterized by (T, U). Understanding the training process of U channel matrix can be complex so we defer the responsibility of this algorithm following [7].

Given the channel matrix U and the statistics f_{sk} and n_{sk} for a segment s, we can perform a class-independent maximum-a- posteriori(MAP) point-estimate of the channel factors x_s, relative to the UBM. This estimate is computed as:

$$\hat{x}_s = (I + \sum_k n_{sk} U_k' U_k) U' f_s. \quad (6)$$

The effect of the channel factors can be approximately removed from the first-order statistic thus:

$$\hat{f}_{sk} = f_{sk} - n_{sk} U_k \hat{x}_s, \quad (7)$$

where $\hat{f}_{sk}$ is the compensated first-order statistic. To get the score, we use the compensated first-order statistic to calculate the class locations:

$$\hat{t}_{ck} = \frac{\sum_s \hat{f}_{sk}}{r + \sum_s n_{sk}}, \quad (8)$$

where r is the relevance factor ($r = 14$ in our experiments). Again, we pack the location supervectors into columns of a matrix denoted as $\hat{T}$ and we score thus:

$$\vec{\lambda}_s = \frac{\hat{T}' \hat{f}_s}{\sum_{k=1}^{1024} n_{sk}}. \quad (9)$$

The scored segments are the ones extracted by using BIC as segmentation algorithm. After this scoring, we use a Gaussian back-end with one Gaussian per class (two dimension Gaussian back-end) as the one described in [8]. A block diagram of the system is presented in Fig 2.

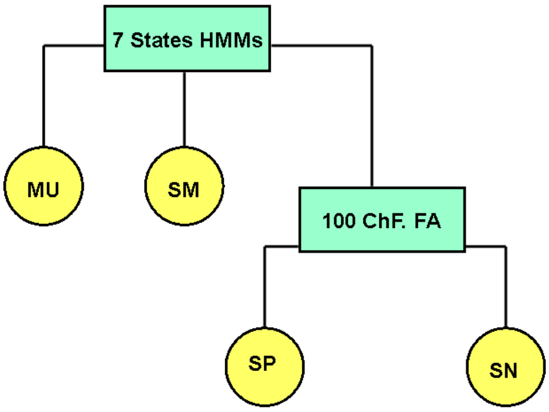

Figure 2: *Block Diagram*

4. Experimental Results

To evaluate the influence of the inclusion of a FA final stage in an HMM/GMM based audio segmentation system, we present first the results of a hierarchical system based only on HMM/GMM. The system configuration is similar to the one that achieved the best performance in the Albayzin 2010 evaluation. The segmentation and classification is performed by using a 7 state HMM with observation probabilities based on GMMs.

Table 1: Segmentation and Classification error for different number of components for the GMMs

	MU	SM	SN	SP	Avg.
32 G	14.9%	30.6%	43.7%	45.3%	33.6%
64 G	16.9%	29.1%	41.2%	43.1%	32.6%
128 G	15.7%	27.8%	41.6%	43.7%	32.2%

Table 1 shows the error rates for different number of components in the GMMs, where only slightly differences can be seen among the different configurations evaluated. As it can be seen the higher error rates are related to two classes, speech and speech with noise, while music and speech with music have error rates significantly lower.

Thus, we focus on reducing the error rates on those classes that present the higher error rates, that is speech and speech with noise, by adding a final FA stage at the end of the classification system instead of the HMM/GMM. We have verified that the use of the FA approach for classification allows a significant reduction of classification errors. This verification has been performed by running experiments where perfect segmentation was considered and only the classification task was evaluated. Classification error rates as low as 19% were obtained by using FA approaches with perfect segmentation while error rates around 30% were obtained by using other approaches as GMM. For the

FA classification system, the UBM and the channel matrix were obtained using all the data available in the training set.

Nevertheless, there is still a challenging problem to be solved for the proposed task, that is, how to integrate the FA approach into a reliable and accurate segmentation system. Among the segmentation systems that we have evaluated, one of the best performance was obtained by using BIC to segment the audio stream and then classify the resulting segments by using FA. In Table 2 the results obtained with this approach are shown along with the ones obtained with the HMM/GMM system. As it can be seen a significant reduction in the error rates for speech and speech with noise has been achieved by using the FA approach. Nevertheless, there is still a long way to run since most of the segmentation and classification error is mainly due to the lack of accuracy of the segmentation system used, that is, BIC to find the borders of the segments.

Table 2: Segmentation and classification error for the proposed hybrid system based on HMM/GMM and Factor analysis compared to results of a system that uses only HMM/GMM.

	MU	SM	SN	SP	Avg.
HMM	15,7%	27,8%	41,6%	43,7%	32,2%
FA	15,7%	27,8%	37,6%	35,4%	29,1%

Finally, Figure 3 shows a comparison among different segmentation and classification systems. The results obtained by the best system that participated in the Albayzin 2010 evaluation [5] are plotted along with the ones obtained with this hybrid architecture that uses both HMM/GMM and FA. The results obtained with two different HMM/GMM systems are also included, the ones obtained with a hierarchical architecture and the ones obtained by not using a hierarchical structure. As it can be seen the best performance is obtained by the HMM/FA hybrid structure for almost every class. The most significant improvement introduced by FA with regard to the HMM/GMM systems, being them hierarchical or not, is in the classification of speech and speech with noise. For example, with HMM/GMM a classification error rate of 43.7% is obtained for the speech class whereas with the FA approach this error rate decreases up to 35.4%. For the speech with noise class, the reduction is from 41.6% to 37.6%.

5. Conclusions

In this work we address the problem of segmentation and classification of broadcast audio. 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 makes use of a final FA stage instead of the classical HMM/GMM to classify between speech and speech with noise, because those were the classes that obtained the higher error rates. Experimental results show that the inclusion of the hybrid structure HMM/FA allows a significant reduction in the classification between speech and speech with noise and thus a reduction in the average segmentation and classification error rate.

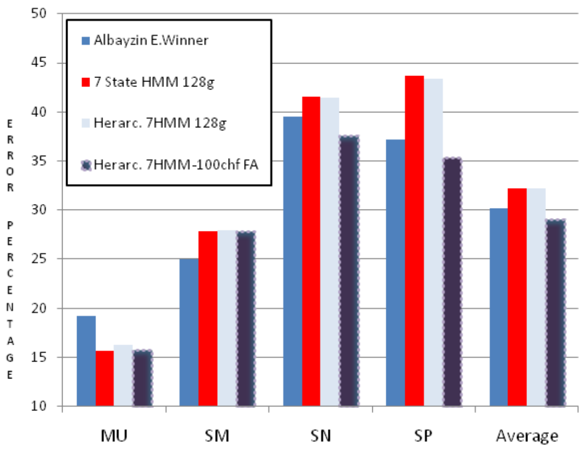

Figure 3: *Performance comparison among different systems*

6. Acknowledgements

This work has been partly funded by the national project TIN2008-06856-C05-04.

7. References

[1] Lu L, Zhang H-J, Li SZ. “Content-based audio classification and segmentation by using support vector machines”. Multimedia Systems. 2003

[2] Nwe TL, Li H. Broadcast news segmentation by audio type analysis. In: Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP05). IEEE International Conference on.Vol 2. IEEE; 2005

[3] Lavner, Y and Ruinskiy, D, A “Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation”, EURASIP Journal on Audio, Speech, and Music Processing, 2009.

[4] Butko T, Nadeu C, Schulz H. Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results. Evaluation. 2010

[5] Gallardo A, San Segundo R. UPM-UC3M system for music and speech segmentation. In: Proc. FALA.; 2010

[6] Brummer N, Strasheim A, Hubeika V, Matejka P, Burget L, and Glembek O. Discriminative acoustic language recognition via channel-compensated GMM statistics. In: Tenth Annual Conference of the International Speech Communication Association.; 2009

[7] Kenny P, Boulianne G, Ouellet P, Dumouchel P. Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. IEEE Transactions on Audio, Speech and Language Processing. 2007

[8] Jancik, Z., Plchot, O., Brummer, N., Burget, L., Glembek, O., Hubeika, V., Karafiat, M., Matejka, P., Mikolov, T., Strasheim, A., Cernocky, J.: Data selection and calibration issues in automatic language recognition - investigation with BUT-AGNITIO NIST LRE 2009 system, In: Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop, Brno, CZ, ISCA, 2010, p. 215-221