# Factor Analysis Segmentation and Classification in Broadcast News Domain

Diego Castan, Alfonso Ortega, and Eduardo Lleida

ViVoLab
Aragon Institute of Engineering Research (I3A)
University of Zaragoza
[dcastan,ortega,lleida]@unizar.es,
WWW home page: http://www.vivolab.es/

**Abstract.** This paper proposes a study of a Factor Analysis (FA) segmentation and classification system. Our approach is inspired by language recognition systems where every input sequence is a language. Following this idea, a study between the classic segmentation systems based on HMM/GMM and FA is done over the output of a perfect segmentation system (oracle boundaries). It can be seen how FA improves the classification results compared to HMM/GMM. Also, the first experiments of an on-building FA segmentation system are reported suggesting the need to improve the channel compensation over some classes.

**Index Terms**: Factor Analysis, Channel Compensation, Broadcast News Segmentation

## 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.

Several approaches have been proposed for audio segmentation in different scenarios. For example, in the task of automatic transcriptions of broadcast news [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 broadcast news 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 hour show with no prior knowledge of acoustic conditions and speakers.

In the context of the Albayzin-2010 evaluation campaign an audio segmentation task was proposed in [6]. Almost all the participants of the evaluation used hierarchical systems, including the winning system [7] based on a hierarchical arquitecture that used different sets of features for every level. For this evaluation database, in [8] 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, a comparation between Factor Analysis and HMM-GMM is reported. The first group of experiments is based on the classification task over the segments with oracle boundaries. In the second group of experiments, the systems must identify the begining and the end of each segment so a segmentation/classification error is reported.

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 language recognition systems. Classifications and segmentation results are presented in section 4. Finally, the conclusions and the future work are presented in section 5.

## 2 Albayzin 2010 audio segmentation evaluation

### 2.1 Database

The database used for the Albayzin2010 evaluation 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 approximately 87 hours of annotated audio (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):

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

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 (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%). Detection error rates (DET) curves are also provided in the hierarchical segmentation systems for comparison.

## 3 Factor Analysis Framework

### 3.1 Statistics

The Factor Analysis approach has been successfully used in speaker recognition [9] and more recently in language recognition [10]. The main advantage of Factor Analysis compared to other classification methods is its ability to compensate for the session variability that can be found in the data due to several factors like background noise, recording devices, etc.

As in language identification, this work examines the problem of assigning a class label to each segment using FA models trying to compensate the within-class variability. Additionally, this task has to deal the problem of detecting boundaries between segments of different classes where every segment may have a different length. These segments are going to be mapped to sufficient statistics of fixed size by using a Universal Background Model (UBM) which is a class-independent GMM trained with the EM-algorithm on the 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 $P_k$ 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 frames indexed $i = 1, 2, ..., N_s$, we define the zero and first-order statistics respectively as:

$$n_{sk} = \sum_{i=1}^{N_s} P_{ksi} \tag{2}$$

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

For 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 assume the UBM as having zero mean and unity precision for all components. After this transformation the formulas below no longer require UBM parameters.

### 3.2 Channel compensation

Data from a particular class segment is 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:

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

where $c(s)$ denotes the class of segment s; $t_{sk}$ is the channel independent class location vector; $U_k$ 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 $U_k$ . 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 equation 4 can be expressed more compactly as:

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

where $U$ is known as the channel matrix and it represents the within-class variability. Let $T = [t_{mu}, t_{ot}, t_{sm}, t_{sn}, t_{sp}]$ where $T$ represents the locations of classes in the GMM space, so our metamodel for class-segment-dependent GMM is parametrized by $(T, U)$ which are describing prior distribution of parameters $m$.

The parameters $\Theta = \{T, U\}$ 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. Inthe E-step posterior distributions of $x$ are estimated for each segment, using current parameters $\Theta_{old}$. In the M-step we find parameters $\Theta$ that maximize the auxiliary function $Q(\Theta, \Theta_{old})$. The simple case is considered where location vectors $t_{ck}$ are obtained by using a single iteration of relevance-MAP adaptation from the UBM. This adaptation is expressed in terms of statistics as:

$$t_{ck} = \frac{\sum_s f_{sk}}{r + \sum_s n_{sk}} \tag{6}$$

where the sums are over all segments s belonging to the class $c$ and $r$ is the relevance factor ($r = 14$ in our experiments). With the class locations fixed, $x$ is re-estimated for each segment $s$ and then $U_k$ for every component $k$.

Given the channel matrix $U$ and the statistics $f_{sk}$ and $n_{sk}$ for a segment s, a class-independent maximum-a- posteriori(MAP) point-estimate of the channel factors $x_s$ can be performed, relative to the UBM as it can be seen in [10]. This estimate is computed as:

$$\hat{x}_s = \big(I + \sum_k n_{sk} U_k' U_k\big)^{-1} U_k' f_s \tag{7}$$

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

$$\hat{f}_{sk} = f_{sk} - n_{sk} U_k \hat{x}_s \tag{8}$$

where $\hat{f}_{sk}$ is the compensated first-orden statistic.

### 3.3 Scoring

In [11], different scoring methods are studied. The log-likelihood ratio (LLR) scoring shows a significant speedup without any loss in performance due to the simplification of scoring shown in [9] by omitting non-linear terms. To get the score, the compensated first-orden statistics are used to calculate the class locations :

$$\hat{t}_{ck} = \frac{\sum_s \hat{f}_{sk}}{r + \sum_s n_{sk}} \tag{9}$$

Again, the location supervectors are packed into the columns of a matrix denoted as $\hat{T}$ and thus the score is computed as:

$$\boldsymbol{\lambda}_s = \frac{\hat{T}' \hat{f}_{sk}}{\sum_{k=1}^{1024} n_{sk}} \tag{10}$$

This type of scoring can be seen as a dot product between the compensated test vector and the different class vectors. As a result, a calibration for the dot product is needed. In our approach, a normal distribution $N(\mu, \Sigma)$ (one Gaussian) is trained using the set of scores vector where each class is represented by one dimension of the Gaussian. This Gaussian transforms the general scores to $N_s$ multi-class log-likelihoods where $N_s$ is the number of target classes [12].

## 4 Experimental results

### 4.1 Factor Analysis as a classifier of segments with oracle boundaries

To evaluate the benefits of using FA, a baseline using the same configuration of the winning system in the Albayzin 2010 evaluation [7] is presented over the output of a perfect segmentation system to be able to evaluate the classification error. This system uses a hierarchical HMM/GMM approach to classify the frames between *MU/NOMU* on the first level, *SM/NOSM* on the second level and *SP/SN* on the last level as it is shown in Fig. 1. The audio features extracted for this system is a combination of 15MFCC + C0 + $\Delta$ + $\Delta\Delta$ + 12Chroma. The mean and the standard deviation are computed over 1 a second window with an overlap of 0.5 seconds. Previous experiments showed us that it is better to use less components in the models of the classes with less data (SM and MU) and more Gaussians per HMM state for the classes with more data (SP and SN). Table 1 shows the average error of the four classes for a different number of states of the HMM and a different number of Gaussians. Note that the number of Gaussians for the SP/SN classes is four times greater than that for the MU/SM classes.

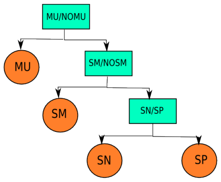


**Fig. 1.** Block diagraman of the hierarchical system

With the same audio features used for the HMM/GMM experiments, a UBM with 1024 Gaussians was trained over all the training set. The channel compensation was performed with 100 channel factors. In table 2 the average error of the four classes over the training dataset, over the test dataset and over the test dataset with a GBE calibration stage are shown. The first row shows the results of the FA over the smoothed features with mean and standard deviation. The use of the mean and the standard deviation for FA seems to be not a good

**Table 1.** Classification error for oracle boundaries with HMM-GMM systems

| Gauss / States | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| 32G-128G | 26.17 | 25.87 | 26.58 | 25.11 | 26.40 | 26.20 | 28.13 |
| 64G-256G | 26.30 | 25.29 | **24.56** | 25.29 | 26.34 | 25.97 | 27.05 |

solution based on the results of table 2 where a good accuracy over the training segments can be seen but a poor generalization over the test segments.

**Table 2.** Classification error for oracle boundaries with FA systems using MFCCs and Chroma features

| | TRAIN | TEST |
|---|---|---|
| With Mean-Std - UBM 1024G - 100ChnF | 3.24 | 55.21 |
| Without Mean-Std - UBM 1024G - 100ChnF | 14.77 | **22.91** |

A very importan issue for the channel compensation task is the early fusion of various types of features. According to the results shown in table 3 stacking features seems not to be the best solution if we compare these figures with the results shown in Table 2.

**Table 3.** Classification error for oracle boundaries with FA systems using MFCCs

| | TEST |
|---|---|
| MFCC16+$\Delta$+$\Delta\Delta$ - UBM 1024G - 100ChnF | 21.25 |
| MFCC16+$\Delta$+$\Delta\Delta$ - UBM 2048G - 100ChnF | **20.81** |

Fig. 2 shows the error per class and the average error for those systems that have better results in each table. The effectiveness of the HMM/GMM system as a music detector compared with FA systems is evident. A possible explanation for this behavior is that the $U$ matrix was trained for all the classes and the most important channel effect is the speech class because it represents 92% of the database so when channel compensation is applied over the music segments, a distortion is produced. On the other hand, the behavior of the FA in SN and SP classes is much better than HMM/GMM.

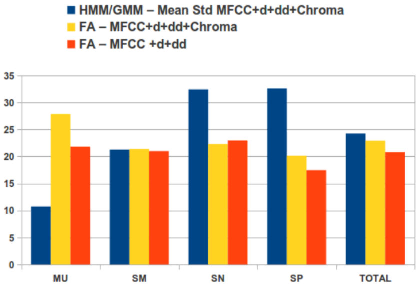


**Fig. 2.** Comparation of error rate per class between HMM/GMM and FA systems

### 4.2 Factor Analysis as a segmentation system

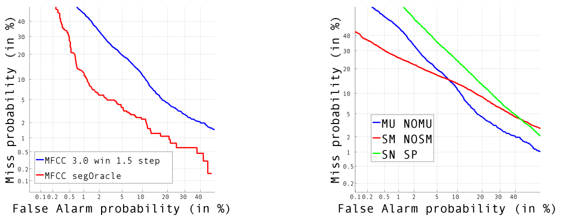
Using the same HMM/GMM hierarchical system with the same audio features that were used in the previous subsection, the error segmentation for different number of states were calculated and the results are presented in table 4. In this case, it is clear that the best configuration for the segmentation task is with 8 HMM states instead of 6 states in the case of the classification task.

**Table 4.** Classification error for oracle boundaries with HMM-GMM systems

| Gauss / States | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| 32G-128G | 39.97 | 34.76 | 32.88 | 31.47 | 30.85 | 30.43 | 31.31 |
| 64G-256G | 39.27 | 33.66 | 31.11 | 30.91 | 30.99 | **29.37** | 31.59 |

For the FA segmentation system we use the MFCC16+$\Delta$+$\Delta\Delta$ audio features with 2048 Gaussians to train the UBM and 100 channel factor to model the channel compensation. As it can be seen, this system was used in the previous subsection yielding the best results in the classification task. The segmentation is produced with the classification of 3 seconds segments with an overlap of 1.5 seconds. In this case the transition probabilities between segments are not used so there is no contextual information at all. With this framework, the difference between the classification error and the segmentation error in the worst case (MU-NOMU because we have more error rate) is evident and is shown in Fig. 3(a) where DET curves for oracle and non-oracle segmentation are compared. In Fig. 3(b) the DET curves for every branch of the hierarchical system are plotted.

Table 5 shows the results plotted in Fig. 3(b) with the evaluation metric for every class. The error for classes like MU or SM is still very high compared to the HMM/GMM error rate. Nevertheless, Fig. 4 shows the DET curves divided



**(a)** **(b)**

**Fig. 3.** (a) DET curves for oracle boundaries vs non-perfect segmentation in the music hierarchical level and (b) DET curves for every level of the hierarchical system

by the length of each segment. It can be seen that for long segments, the GMM is the best classifier but for short segments, FA system is much better than the GMM.

**Table 5.** Segmentation error per class for the best HMM/GMM system and FA system

| | MU | SM | SN | SP | TOTAL |
|---|---|---|---|---|---|
| HMM/GMM-8states | 15.93 | 23.43 | 38.66 | 39.48 | 29.37 |
| Hierarchical FA | 52.91 | 37.19 | 45.08 | 40.80 | 43.99 |

## 5 Conclusion and future work

By means of classification experiments it has been shown that channel compensation helps to classify segments decreasng the error rate and improving the classification of all speech classes. These results justify the creation of a whole-FA segmentation system following the same hierarchical structure used for HMM/GMM. Although the segmentation error is very high, the better classification in short segments encourages to improve the system.

For future work, different window lengths and time advances will be implemented to try to improve the segmentation. Also, other scoring methods different than the linear scoring will be studied like those presented in [11]. In addition, the class dependent training of several $U$ matrices will be investigated creating a new $U$ matrix by stacking the different class dependent $U$ matrices to decrease the error in the MU class which is critical for the metric of the evaluation.

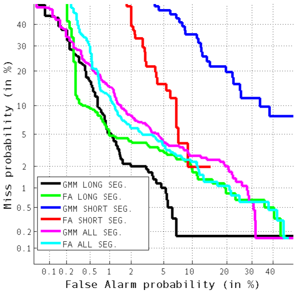


**Fig. 4.** DET curves GMM vs FA with different length of segments

## References

[1] Chen SS, Gopalakrishnan PS. “IBM L CSR System for Transcription of Broadcast News Used in the 1997 Hub4 English Evaluation”. Proceedings of the Speech Recognition Workshop, 1998.

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

[3] 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

[4] 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.

[5] Siegler MA, Jain U, Raj B, Stern RM. “Automatic Segmentation, Classification and Clustering of Broadcast News Audio”. Signal Processing.4-6.

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

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

[8] Castan D, Vaquero C, Ortega A, Martinez D, Lleida E. “Hierarchical Audio Segmentation with HMM and Factor Analysis in Broadcast News Domain”. In: Interspeech.; 2011.

[9] 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

[10] Brummer N, Strasheim A, Hubeika V, et al. “Discriminative acoustic language recognition via channel-compensated GMM statistics”. In: Tenth Annual Conference of the International Speech Communication Association.; 2009

[11] Glembek O, Burget L, Dehak N, Brummer N, Kenny P. “Comparison of scoring methods used in speaker recognition with Joint Factor Analysis”. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing; 2009.

[12] Benzeghiba MF, Gauvain J-luc, Lamel L, Cnrs L, Cedex BPO. “Language Score Calibration using Adapted Gaussian Back-end”. In: Interspeech 2009.