A preliminary study of Acoustic Events Classification With Factor Analysis in Meeting Rooms

Diego Castán, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida

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

Abstract. The classification of acoustic events is useful to describe the scene and can contribute to improve the robustness of different speech technologies. However, the events are usually overlapped with speech or other sounds. This work proposes an approach based on Factor Analysis to compensate the variability of the acoustic events due to overlap with speech. The system is evaluated in the CLEAR evaluation database composed of recordings in meeting rooms where the acoustic events have been spontaneously generated in five different locations. The experiments are divided in two sets. Firstly, isolated acoustic events are used as development to analyze and evaluate parameters of the Factor Analysis system. Secondly, the system is compared to a baseline based on Gaussians Mixture Models with Hidden Markov Models. The Factor Analysis approach improves the total error rate due to the variability compensation of overlapped segments.

Index Terms: Acoustic Events, Factor Analysis, Meeting Rooms, CLEAR Evaluation

1 Introduction

Speech can be considered the most informative part of the audio. However, non-speech sounds can be useful to characterize situations of people, places or objects. These sounds are known in the literature as *acoustic events* (AEs) and can be critical to understand human activities or to describe the scene. *Acoustic Event Detection* (AED) aims at processing a continuous audio stream and determine what event has been produced and when. Therefore, the system must be able to produce labels to understand the concept behind the event. The symbolic description provided by the AED systems has been used in a wide variety of applications. For example, in [1] the authors detect events for surveillance. AED has been also widely used for monitoring people with disabilities. An example of

this is given in [2] where people with dementia can be monitored in the bathroom to generate an automatic hygiene behavioral report.

AED is very useful in the task of audio indexing and retrieval of multimedia documents or related tasks like multimedia event detection (MED) since it is an important resource of semantic description. For example, in [3], the authors combine the AEs with speech to detect five different multimedia events. Therefore, some approaches in this field try to compensate the variability with factor analysis (FA) techniques [4] or model the temporal relationship between events [5] to provide robustness to the AED system.

Another challenging field is the AED in meeting rooms because the AEs have low SNR and are overlapped with speech or other AEs. The 2007 AED CLEAR Evaluation [6] was performed on a database recorded in real seminars (five different locations) where the AEs were spontaneously generated, most of them are not highlighted and overlapped with speech. In this evaluation, the submitted systems showed low accuracies and high error rates (the winning system [7] got around 30% of accuracy and 99% of error rate) where the overlapping segments represent more than 70% of the errors. Subsequent investigations have dealt with the overlap problem in different ways. Since the meeting rooms in the evaluation are equipped with multiple cameras and multiple microphone arrays, in [8] the authors propose a multi-modal system because some of the AEs have a visual correlate and, therefore, the video modality can be exploited to enhance the detection rate. Also the authors use multiple microphones to know the position of the AE since some events can only occur at particular locations like "door slam". Another popular solution is the separation of overlapped signals with signal processing techniques. In [9], an approach based on partial signal separation using multiple array beamformers was proposed previously to an HMM-GMM classification system.

Since the CLEAR evaluation database was recorded in five different rooms with different furniture, the AEs present some variability that can be compensated. This work studies variability compensation techniques based on factor analysis with one microphone in this meeting room environment. The main goal is to increase the robustness in the classification of the AEs for each microphone so it does not interfere with multimodal or multichannel techniques that could be applied later. Due to the extremely high error rate shown in the CLEAR evaluation, this paper proposes a preliminary study where the segmentation is given by the labels to evaluate the classification of the proposed system and leaving the detection as a future work.

The remainder of the paper is organized as follows: section 2 describes the database and the metric for this task. The FA framework is described in section 3. Section 4 shows a comparative of the proposed system with a baseline and, finally, the conclusions are presented in section 5.

2 Database and Metric

2.1 Database

The database used in the CLEAR evaluation is composed of 25 meetings of approximately 30 minute long recorded in five different meeting rooms (AIT, ITC, IBM, UKA and UPC). One meeting from each location have been used as training set and the rest of the meetings represents the test set. Also, a database with isolated AEs recorded at UPC [10] has been used to get some preliminary results, but these isolated events have not been used to train the final system.

The set of AEs composed of 12 semantic classes is summarized in Table 1. The classes of "speech", "unknown" and "silence" are not evaluated. 11% of the database is silence, 53% are "speech" and "unknown" classes and 36% of the time are AEs where most of them are overlapped with "speech" (64%) or other AEs (3%).

Table 1: Acoustic event classes with the corresponding annotation label and the number of the events in the train and test set

Event name	**Label**	**Train**	**Test**
Knoch in door or table	[kn]	82	152
Door slam	[ds]	73	75
Step	[st]	72	496
Chair moving	[cm]	238	226
Cup jingle	[cl]	28	27
Paper wrapping	[pw]	130	88
Key jingle	[kn]	22	32
Keybord typing	[kt]	72	105
Phone ringing or music	[pr]	21	25
Applause	[ap]	8	13
Cough	[co]	54	36
Laugh	[la]	37	154
Unknown (Unidentified sounds)	[un]	-	-
Speech	[sp]	-	-
Silence	[ ]	-	-

2.2 Classification Metric

In these experiments, the system has to classify correctly the segment which boundaries are given by the reference labels. The segments where two AEs are overlapped count twice (one for each event). The error rate for the acoustic event classification (AEC-ER) can be written as:

$$AEC - ER = \frac{number\ of\ segments\ incorrectly\ classified}{number\ of\ total\ segments} \quad (1)$$

3 Factor Analysis Framework

We propose a framework for AED system that deals with the problem of assigning a class label to each fixed-length window using Factor Analysis (FA) models. The FA approach has been successfully used in speaker recognition/verification, speaker diarization and in language recognition. In these tasks, the systems have to face several sources of variability such as speaker, channel and environment. The variability of the same class segments is known as *within-class variability*. The goal of these systems is to model or compensate the *within-class variability* to reduce the mismatch between training and test. In our task, the variability comes from the overlap of the AEs with speech and the different locations where the database has been recorded. The functionality of this system is described in the next subsections step by step.

3.1 Acoustic feature extraction

In this work we extract 16 MFCCs (including the zeroth order cepstrum) computed in 25 ms frames with a 10 ms frame step, their first and second derivatives. The feature vectors are normalized in mean for each file.

3.2 Channel compensation

A particular class is modeled by a GMM defined by a set of mean vectors $\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_C$, weights $w_1, w_2, ..., w_C$ and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\Sigma}_C$ where C is the number of Gaussians. We can concatenate all GMM mean-vectors to one mean supervector $\boldsymbol{m}$ of dimension $CF \times 1$ where F is the feature vector length:

$$\boldsymbol{m} = [\boldsymbol{m}_1^T, \boldsymbol{m}_2^T, ..., \boldsymbol{m}_C^T]^T. \tag{2}$$

The Factor Analysis model is the adaptation of a general GMM model (known as Universal Background Model or UBM in the literature) where the supervector of means is not fixed and it can vary from segment to segment due to several sources that increase the within-class variability. We assume that these GMMs have segment and class dependent means but fixed weights and covariances chosen to be equal to the UBM weights and covariances. Specifically, we use a Factor Analysis model for the mean vector of the kth component of the GMM for segment s:

$$\boldsymbol{m}^s = \boldsymbol{t}^{c(s)} + \boldsymbol{U}\boldsymbol{x}_s, \tag{3}$$

where $c(s)$ denotes the class of segment s. The class-location vector $\boldsymbol{t}^{c(s)}$ is obtained by using a single iteration of relevance-MAP adaptation from the UBM. $\boldsymbol{U}$ is known as the *within-class variability matrix* and $\boldsymbol{x}_s$ is a vector of L *segment-dependent-within-class-variability factors* assumed to follow a normal distribution ($N(0, I_L)$). The columns of the $\boldsymbol{U}$ matrix are the basis spanning the subspace of the within-class variability and the *within-class variability factors* are

the coordinates defining the position of the supervector in the subspace. The *within-class variability factor* dimension (L) is smaller than CF so $\boldsymbol{U}$ has low rank ($CF \times L$ dimensions). This paper does not aim to deepen in the FA theory and more details with an exhaustive description can be found in [11].

3.3 Class/non-class models

Most of the approaches based on FA are implemented with a single $\boldsymbol{U}$ matrix because the segments are well-delimited (typically in separate files) and the nature of the within-class variability is similar for all the classes. However, in [12], an approach based on FA was proposed with class/non-class vectors (one class vector and one non-class vector for each class) and specific matrices modeling the within-class variability of each pair class/non-class as follows:

$$\boldsymbol{T} = [\boldsymbol{t}^{class}, \boldsymbol{t}^{\overline{class}}] \tag{4}$$

$$\boldsymbol{U} = \boldsymbol{U}^{class-\overline{class}} \tag{5}$$

The main advantage of these models is that the final scoring can be more discriminative. For example, in the speaker ID tasks, the score to detect a speaker is the log-likelihood ratio test (LLRT):

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

where the numerator is the likelihood for the class model and the denominator the likelihood for the UBM. Note that the UBM is used as a general model to describe the alternative hypothesis which is appropriate for speaker identification where the hypothesized speaker is not in the UBM. However, in a problem with a small number of classes, a non-class model can be trained to be used as the alternative hypothesis as:

$$CLLRT_{class} = log\frac{P(\chi/class)}{P(\chi/\overline{class})}, \tag{7}$$

where the alternative hypothesis is the likelihood for the non-class model which is compensated also with the with-in class variability matrix.

4 Experimental Results

Two different set of experiments have been carried out with a clear increase in the difficulty of the task. The first set is composed of isolated AEs with oracle boundaries and the second set verifies the quality of the models to classify the overlapped AEs when the boundaries are given. Only one microphone located in the center of the room (on the table) was used for both experiments to be able to capture all the activity of the meeting.

4.1 Classification of Isolated Acoustic Events

The AEs used in this experiments were recorded in the UPC smart-room for development. Although the AEs are the same than the CLEAR AEs shown in Table 1, these isolated AEs are not used in the posterior experiments because the AEs are not generated in an spontaneous way and they are not overlapped with speech or other AEs. However, this experiment is useful to study the behavior of the proposed system, to set some parameters and it shows how the errors are distributed among different AEs. The database is divided into three groups: two of them are used to train the model and the third one is used to test.

A GMM with 128 components has been used as a baseline system. Table 2 shows the error rate given by the metric of eq.(1). The same experiment has been done with FA where the UBM was also trained with 128 Gaussians over all the train set to be able to compare the results with the baseline. Table 2 shows the results of this experiment with the baseline and with FA for different values of τ. This parameter is known as *relevance factor* (τ) and it controls the MAP adaptation of the means of the model. If we increase the τ to infinite, the MAP will remain in the original UBM. On the other hand, if we decrease τ, the means will be more affected by the new frames. The results show that it is better to be more restrictive in movements of the means and, therefore, a $\tau = 250$ is chosen for the next experiments.

Table 2: Classification error rate for CHIL acoustic events with oracle segmentation

System	**Error Rate %**
GMM-128G	3.26
FA $\tau = 10$	5.22
FA $\tau = 30$	4.47
FA $\tau = 50$	4.24
FA $\tau = 100$	4.24
FA $\tau = 250$	3.26
FA $\tau = 500$	3.26

The error rate is equal with GMM and FA and no relevant conclusions can be drawn. However, this experiment allows to chose the τ parameter for next experiments with overlapped AEs. Also, Figure 1 shows the error of the classes with GMM and FA for different values of τ and, as it can be seen, only half of the AEs are not correctly classified once or more. Also, increasing τ, the classification accuracy improves. Finally, some AEs are better classified with GMM ([pw] and [pr]) and others are better classified with FA ([co] and [st]). We can conclude that both systems classify the isolated events easily because these AEs are artificially generated and there is not overlap with speech or other events. Therefore, the variability of each AE is very reduced and both systems classify with the same accuracy.

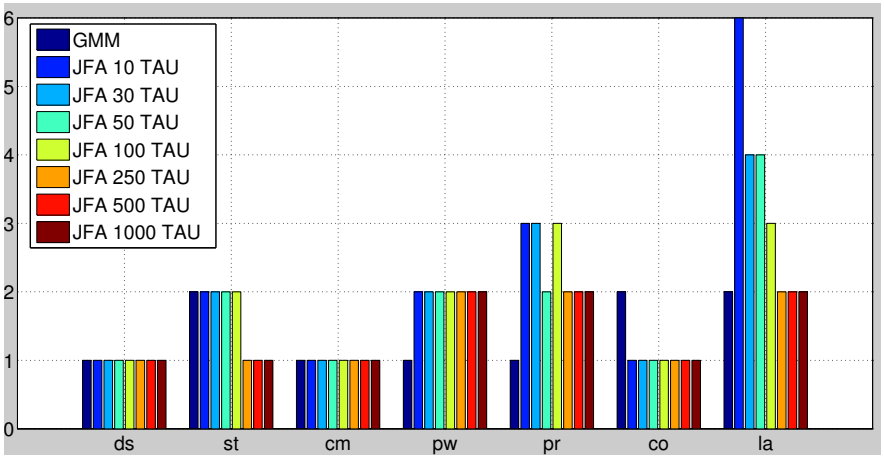


Fig. 1: Number of errors for each acoustic event

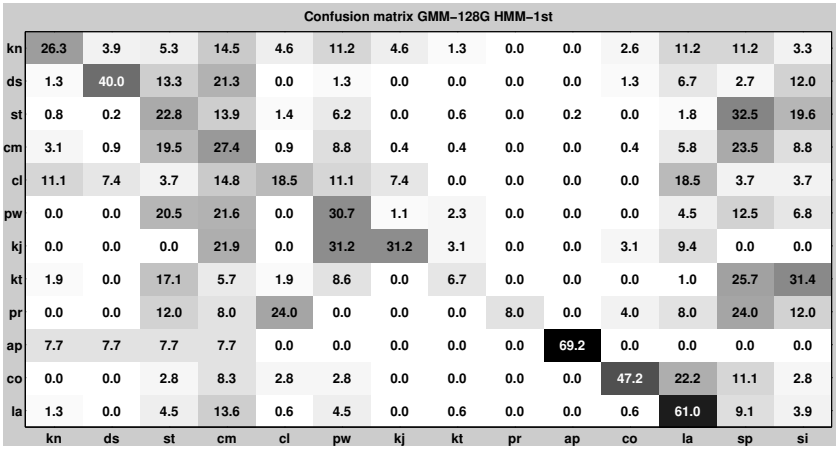
4.2 Classification of Spontaneous Acoustic Events

Following the procedure presented in the last subsection, this experiment is carried out with oracle segmentation over the CHIL database where the events can be overlapped with speech or other events increasing the difficulty dramatically. In addition, the audio has been recorded in five different locations which increase the variability of each event.

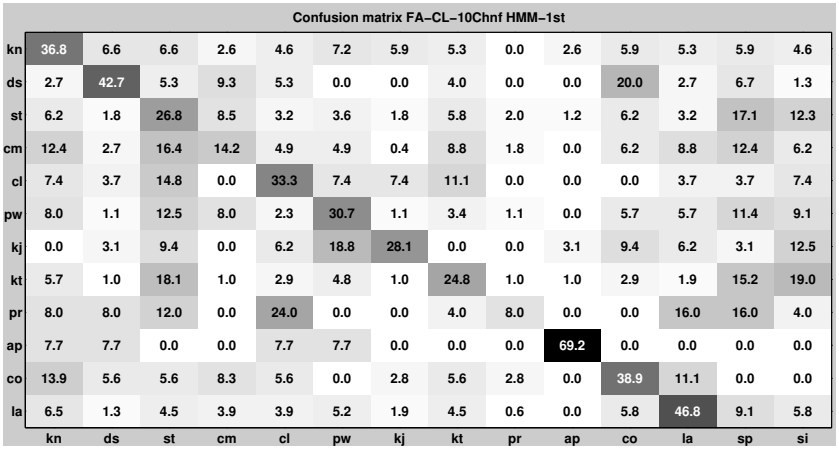
Table 3: Classification error rate for CHIL acoustic events with oracle segmentation

System	**Error Rate %**
GMM-128G	70.95
GMM-128G / one state HMM	70.88
FA-CLNoCL-10Chnf	75.71
FA-CLNoCL-10Chnf / one state HMM	75.57
FA-CL-10Chnf	71.09
FA-CL-10Chnf / one state HMM	**70.11**

Table 3 compares a baseline based on GMM, with FA system. The parameters for this experiment are the values fixed in the last subsection: 128 Gaussian for GMMs and UBM and $\tau = 250$ for the MAP adaptation. The first two rows show the error rate for a GMM-128G and the same GMM inside a one-state HMM where the transition probabilities have been estimated with the training labels improving slightly the results compared to the GMM system. The remaining rows show the error rate for the FA systems with two different scores: the FA-CLNoCL approaches employ eq.7 while the FA-CL approaches employ eq. 6. The results clearly show that the FA-CL approaches are more discriminative than the FA-



(a)



(b)

Fig. 2: Confusion matrices for (a) GMM-128G / HMM-1st and (b) FA-CL-10Chnf / HMM-1st

CLNoCL since the model and the anti-model share common information due to events overlapped with speech. Therefore, the best result is given for the system FA-CL which slightly improves the final result with the transition probabilities compared to the HMM-GMM.

Finally, Figure 2 shows the confusion matrices with the classification percentage in each event combination for the best two systems: GMM-128G with one state HMM approach and the FA-CL with one state HMM approach. Some conclusion can be drawn from these figures. First, the GMM system tends to classify the AEs as “speech” [sp] or “silence” [si] more easily than the FA which shows that the FA system compensates the variability due to the speech in the

overlapped AEs. Also, the FA system classifies better the events "Door Knock" [kn], "Door Slam"[ds], "Steps" [st], "Cup Jingle" [cl] and "Keyboard Typing" [kt]. On the other hand, the events "Chair Moving" [cm], "Key Jingle" [kj], "Cough" [co] and "Laugh" [la] have been better classified with GMM. The rest of the AEs have been classified with identically accuracy. A final count shows that the GMM and the FA have correctly classified 416 and 421 AEs respectively from a total of 1429 AEs.

5 Conclusions

The presented work focuses on the classification of AEs that may happen in a meeting room when the segmentation is given using the CLEAR evaluation database. Since the database is composed of tracks recorded in five different locations and the events can be overlapped with speech, the AEs present a variability that can be compensated with FA techniques. Two sets of experiments have been carried out in this work. The first one evaluates the FA system over isolated AEs. This isolated AEs database has been used as a development to choose the *relevance factor* (τ) of the MAP adaptation for the FA systems. The second set of experiments evaluates the FA system with spontaneous generated AEs that can be overlapped with speech or other AEs. The proposed system improves the results of a baseline system based on GMM/HMM slightly. The confusion matrices of both systems suggest that the FA system compensates the variability due to the speech in the overlapped AEs. However, there is still a big room from improvement since the classification error is very high. Therefore, further work needs to be done to improve the classification of overlapping sounds and the application of FA techniques to the detection problem.

Acknowledgements

This work has been funded by the Spanish Government and the European Union (FEDER) under the project TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

References

1. P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli. Audio Based Event Detection for Multimedia Surveillance. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, pages V–813–V–816, 2006.
2. Jianfeng Chen, Jianmin Zhang, Alvin Harvey Kam, and Louis Shue. An Automatic Acoustic Bathroom Monitoring System. *2005 IEEE International Symposium on Circuits and Systems*, pages 1750–1753, 2005.
3. Julien van Hout, Murat Akbacak, Diego Castan, Eric Yeh, and Michelle Sanchez. Extracting Spoken and Acoustic Concepts For Multimedia Event Detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2–6, 2013.

4. Zhen Huang, You-chi Cheng, Kehuang Li, Ville Hautamaki, and Chin-hui Lee. A Blind Segmentation Approach to Acoustic Event Detection Based on I-Vector. In *Proc. Interspeech*, number August, pages 2282–2286, 2013.
5. Diego Castan and Murat Akbacak. Indexing Multimedia Documents with Acoustic Concept Recognition Lattices. In *Interspeech*, pages 3–7, 2013.
6. A. Temko, Robert Malkin, Christian Zieger, and Dusan Macho. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. In *IV Jornadas en Tecnología del Habla*, pages 1–6, 2006.
7. X Zhou, Xiaodan Zhuang, and Ming Liu. HMM-based acoustic event detection with AdaBoost feature selection. In *CLEAR 2007*, 2008.
8. Taras Butko and C Nadeu Camprubí. Detection of overlapped acoustic events using fusion of audio and video modalities. In *Proc. FALA*, pages 165–168, 2010.
9. Rupayan Chakraborty. *Acoustic Event Detection and Localization using Distributed Microphone Arrays*. PhD thesis, 2013.
10. A. Temko, D. Macho, C. Nadeu, and C. Segura. UPC-TALP Database of Isolated Acoustic Events. In *Internal UPC report*, 2005.
11. P. Kenny, G. Boulianne, Pierre Ouellet, and P. Dumouchel. Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Trans Audio Speech Lang*, 15(4):1435–1447, May 2007.
12. Diego Castan, Alfonso Ortega, Jesus Villalba, Antonio Miguel, and Eduardo Lleida. SEGMENTATION-BY-CLASSIFICATION SYSTEM BASED ON FACTOR ANALYSIS. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.