Improving Detection of Acoustic Events Using Audiovisual Data and Feature Level Fusion

T. Butko[1,2], C. Canton-Ferrer[1], C. Segura[1,2], X. Giró[1], C. Nadeu[1,2], J. Hernando[1,2], J.R. Casas[1]

[1]Department of Signal Theory and Communications
[2]TALP Research Center
Technical University of Catalonia, Barcelona, Spain
{butko, ccanton, csegura, xgiro, climent, javier, josep}@gps.tsc.upc.edu

Abstract

The detection of the acoustic events (AEs) that are naturally produced in a meeting room may help to describe the human and social activity that takes place in it. When applied to spontaneous recordings, the detection of AEs from only audio information shows a large amount of errors, which are mostly due to temporal overlapping of sounds. In this paper, a system to detect and recognize AEs using both audio and video information is presented. A feature-level fusion strategy is used, and the structure of the HMM-GMM based system considers each class separately and uses a one-against-all strategy for training. Experimental AED results with a new and rather spontaneous dataset are presented which show the advantage of the proposed approach.

Index Terms: acoustic event detection, multimodality, multimodal fusion, hidden Markov models, acoustic localization

1. Introduction

In context-aware systems such as smart rooms or intelligent personal devices, Acoustic Event Detection (AED) can provide support for a high-level analysis of the underlying acoustic scene. This analysis includes the description of human activity which is reflected in a rich variety of AEs, either produced by the human body or by objects handled by them. Moreover, AED can contribute to improve the performance and robustness of speech technologies such as speech and speaker recognition, and speech enhancement.

Although speech is usually the most informative AE, other kind of sounds may carry useful cues for scene understanding. For instance, in a meeting/lecture context, we may associate a chair moving or door noise to its start or end, cup clinking to a coffee break, or footsteps to somebody entering or leaving. Furthermore, some of these AEs are tightly coupled with human behaviors or psychological states: coughing or paper wrapping may denote tension; laughing, cheerfulness; yawning in the middle of a lecture, boredom; keyboard typing, distraction from the main activity in a meeting; and clapping during a speech, approval.

AED is usually addressed from an audio perspective and most of the existing contributions are intended for indexing and retrieval of multimedia documents [1] or to improve robustness of speech recognition [2]. Within the context of ambient intelligence, AED applied to give a contextual description of a meeting scenario was pioneered by [3]. Moreover, AED has been adopted as a semantically relevant technology in several international projects [4] and evaluation campaigns [5]. According to results from recent evaluations on AED [5], the single main factor that accounts for the observed low AE detection scores is the high degree of overlap between sounds, especially between the targeted acoustic events and speech. That overlap problem may be faced by developing efficient algorithms that use additional modalities that are less sensitive to the overlap phenomena present in the audio signal.

Most of human produced AEs have a visual correlate that can be exploited to enhance detection and recognition rates. This idea was first presented in [6] where the detection of footsteps was improved by exploiting the velocity information obtained from a video-based person-tracking system. Further improvement has been achieved by the authors in [7] where the concept of multimodal AED is extended to detect and recognize the set of 11 AEs. In that work, not only video information but also acoustic source localization information was considered. A decision-level fuzzy integral fusion was used to increase the accuracy of detection of isolated AEs.

In this paper, we compare that previous approach [7] with a feature-level fusion strategy and present results for AED on new and rather spontaneous data, where the temporal overlaps of sounds take place more frequently. Additionally, a new HMM-GMM based AED system structure is used which considers each class separately and uses a *one-against-all* strategy for training.

Although the above mentioned meeting-room events are no longer acoustic but audio-visual, in this paper we refer to acoustic events, because the audio characterization of events provides the main description for them. The event is considered when it has a specific audio counterpart (sound activity), and video information is only an additional source of information which is used to enhance the audio mono-modal recognition. Actually, in the employed multimodal database, the main criterion for annotating a particular instance of a given class is the existence of acoustic activity.

For this research work, a multi-camera and multi-microphone dataset containing a large number of instances of the AEs to be analyzed has been recorded and manually annotated, and it is available for research purposes[1].

The rest of this paper is organized as follows: Section 2 describes the database and metrics used in evaluations. The baseline detection system is described in Section 3. In Section 4 the fusion approach of different modalities is outlined. Section 5 presents experimental results and Section 6 concludes the work.

2. Database and metrics

There are several publicly available multimodal databases designed to recognize events, activities, and their relationships

[1] Please contact any of the authors for further information

in interaction scenarios [4]. However, these data are not well suited to audiovisual AED since the employed cameras do not provide a close view of the subjects under study. A new database has been recorded with 5 calibrated cameras at a resolution of 768x576 at 25 fps, and 6 T-shaped 4-microphone clusters are also employed, sampling the acoustic signal at 44kHz. Synchronization among all sensors is fulfilled. The database includes two kinds of datasets: recordings of isolated AEs, where 4 different participants performed 10 times each AE, and a more spontaneously generated dataset which consists of 9 scenes of about 5 min long with 2 participants that interact with each other in a natural way: drink coffee, speak on the mobile phone, etc. All AEs appear with a natural frequency: for instance, applause appears much less frequently (1 instance per scene) than chair moving (around 8 instances per scene).

Manual annotation of the data has been done to get a reliable performance evaluation. In order to encourage other researchers to work on this multimodal AED field, these datasets will be made publicly available. The metric defined in [5] is employed to assess the accuracy of the presented algorithms. This metric is defined as the harmonic mean between precision and recall scores computed for the classes of interest.

It must be mentioned that it is difficult to record a large number of AEs in spontaneous gatherings. Indeed, the number of AEs present in these recordings is very low, thus requiring several hours of data. If we force to produce more events during seminar recordings, the resulting AEs are not spontaneous anymore.

3. AED system based only on audio features

A set of spectro-temporal features is extracted to describe every audio frame. It consists of 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives [8], which represent the spectral envelope of the audio waveform within a frame (30ms length, 20ms shift, Hamming window), as well as its temporal evolution. In audio recognition and retrieval, the segments are often modeled via Gaussian Mixture Models (GMMs). An alternative approach presented in [3] exploits discriminative Support Vector Machines (SVM) models to obtain a binary sequence of decisions. In this work, we use HMMs, like in [6]. The topology of the detection process is depicted in Fig. 1.

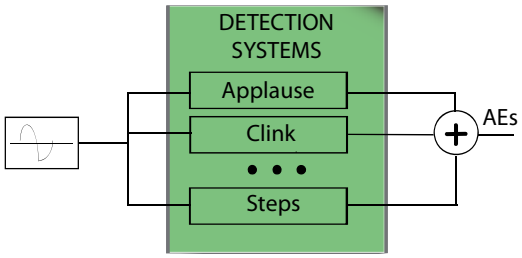


Figure 1: *AED system based only on audio features.*

There are 11 separate systems, each one detecting the targeted class “Class”, so segmenting the audio waveform in intervals labeled as either “Class” or “nonClass”. Using the training approach known as *one-against-all* method [9], all the classes different from “Class” are used to train the “nonClass” model. The model for “Class” is a HMM with 3 emitting states and left-to-right connected state transitions. The observation distributions of the states are Gaussian mixtures with continuous densities, and consist of 5 components with diagonal covariance matrices. The “NoClass” model consists of 5 emitting states with 7 Gaussian components and left-to-right connected state transitions, as its observation distribution is more complex.

The proposed architecture has several advantages:

1. For each particular AE, the best set of features is used. The features which are useful for detection one class are not necessarily useful for other classes. In our case the video features are useful only for detection particular classes.

2. The relation between the number of missings and false alarms can be optimized for each particular AE.

3. In the case of overlapped AEs, the proposed system can provide multiple decisions for the same audio segment.

However this architecture requires multiple detection systems instead of one, which makes the detection process more complex in the case of a large number of classes.

Table 1. *Comparison between baseline systems.*

AEs	Isolated events, accuracy	
	Bas1	Bas2
Applause	0.90	0.89
Cup clink	0.87	0.93
Chair moving	0.89	0.94
Cough	0.78	0.81
Door slam	0.92	0.97
Key jingle	0.92	0.94
Knock	0.91	0.91
Keyboard	0.84	0.96
Phone	0.69	0.87
Paper work	0.74	0.85
Steps	0.18	0.31

3.1. Analysis of results

The comparison of two baseline detection systems for dataset of isolated AEs is presented in table 1. The first system (Bas1) exploits the classical approach, where all AEs are detected inside a unique AED system [7]. The second (Bas2) exploits the *one-against-all* topology described in the previous section. Almost all events are detected better using the second approach. The main reason of such improvement lies in the optimization of the tradeoff between missings and false alarms for each particular class (the word insertion penalty parameter during Viterbi decoding). So this topology is selected for the subsequent experiments with the spontaneous dataset.

Regarding the spontaneous data, the most difficult AEs are low-energy events as keyboard typing, paper work, and steps. Two types of mistakes are associated with these classes:

Type I: Overlap mistakes, when two different AEs occur at the same moment and the detection system fails to detect both of them simultaneously (about 70% of all errors).

Type II: Confusion mistakes. Appear when the classes sound very similar and the AED system fails to distinguish between them (about 30% of all errors)

The overlap problem can be solved at the signal level by means of acoustic source separation techniques or including additional features coming from video modality that are less sensitive to the overlap phenomena. Moreover, we will see that the mistakes of type II can be partially reduced by using features coming from acoustic source localization.

4. AED system based on audio, video and localization features

The overall operation of the proposed system is depicted in Fig.2. First, two information sources correspond to acoustic data processing: single channel audio provides spectro-temporal features, while microphone array processing estimates the 3D location of the audio source. Second, data from multiple cameras covering the scenario allows extracting cues related to some AEs by means of several video-based technologies: person tracking, motion analysis, and object detection. The fusion of modalities is done at the feature level by concatenating features in one super-vector. The final segmentation of the audio waveform is based on a HMM classifier.

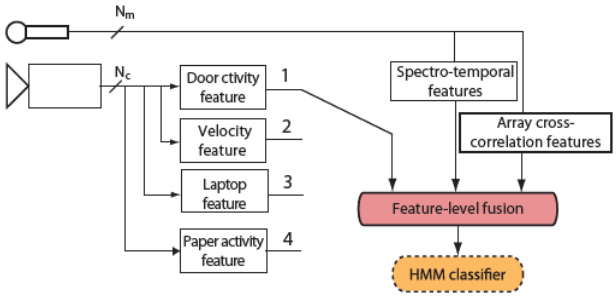
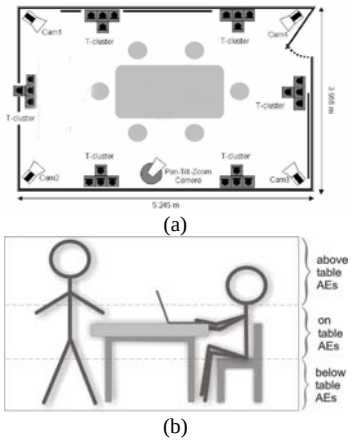


Figure 2: *System flowchart.*

4.1. Room model and localization features

To enhance the recognition results of the baseline system additional features are proposed. In our case, as the characteristics of the room are known beforehand (Fig. 3 (a)), the position *(x, y, z)* of the acoustic source may carry useful information. In fact, events as door slam and door knock can only appear near the door, so a feature which describes the distance from the door is employed in this paper. On the other hand, usually each AE has an associated height, so the *z* position of the acoustic source may help to distinguish among AEs. The following categories are defined as indicated in Fig. 3 (b): *below table, on table, and above table.*



(a)

(b)

Figure 3: *(a) The top view of the room. (b) The three categories along the vertical axis.*

The acoustic localization system used in this work is based on the SRP-PHAT [10] localization method, which is known to perform robustly in most scenarios. In short, this algorithm consists of exploring the 3D space, searching for the maximum of the global contribution of the PHAT-frequency-weighted cross-correlations from all the microphone pairs.

4.2. Video features

Tracking of multiple people present in the analysis area basically produces two figures associated with each target: position and velocity. The human velocity is readily associated to the footsteps AE. Multiple cameras are employed to perform tracking of several people interacting in the scene, by applying the real-time performance algorithm presented in [11].

The motion visual analysis is also used to detect two other acoustic events: paper wrapping and door slam. A motion of a white object near a human in the scene can be associated to paper wrapping (under the assumption that a paper sheet is distinguishable from the background color). The movement of the door can be well detected by the camera oriented towards the door. In order to visually detect a door slam AE, we exploited the a-priori knowledge about the physical location of the door. Analyzing the zenital camera view, activity near the door can be addressed by means of a foreground/background pixel classification [12]. A high enough amount of foreground pixels in the door area will indicate that a person has entered or exited, hence allowing the visual detection of a door slam AE.

Detection of certain objects in the scene can be beneficial to detect AEs such as phone ringing, cup clinking or keyboard typing. Unfortunately, phones and cups are too small to be efficiently detected in the scene but, the case of a laptop can be addressed. In our case, the detection of laptops is performed from a zenital camera located at the ceiling of the scenario. The algorithm initially detects the laptop's screen and keyboard separately and, in a second stage, assesses their relative position and size [13]. Once the position of the laptop is detected, the amount of "skin" pixels over this position will allow to decide about a keyboard typing AE.

4.3. Feature-level fusion approach

Information fusion can be done at different levels: data, feature, and decision level. Data-level fusion is rarely found in multi-modal systems because raw data are usually not compatible among modalities. Concatenating feature vectors from different modalities into one super-vector is a possible way for combining audio and visual information. This approach has been reported in [14] for multimodal speaker recognition.

In this work we use a HMM-GMM approach with feature-level fusion, which is implemented by concatenating the feature sets X_s from S different modalities in one super-vector:

$$Z = X_1 \cup X_2 \cup ... \cup X_S .$$

Then, the likelihood of that observation super-vector at state j and time t is calculated as:

$$b_Z(t) = \sum_m p_m N(Z_t; \mu_m; \Sigma_m),$$

where $N(.;\mu;\Sigma)$ is a multi-variate Gaussian pdf with mean vector μ and covariance matrix Σ, and p_m are the mixture weights. Assuming uncorrelated feature streams, diagonal covariance matrices are considered.

Table 2. *Fusion results*.

AEs	Spontaneous AEs			
	Bas	Bas+L	Bas+V	Bas+L+V
Applause	0.83	0.83	-	-
Cup clink	0.90	0.86	-	-
Chair moving	0.82	0.84	-	-
Cough	0.76	0.82	-	-
Door slam	0.74	0.82	0.85	0.87
Key jingle	0.48	0.39	-	-
Knock	0.86	0.90	-	-
Keyboard	0.71	0.78	0.79	0.80
Phone	0.87	0.90		
Paper work	0.65	0.62	0.73	0.79
Steps	0.58	0.58	0.70	0.66

5. Experiments and results

In order to prove the adequateness of the proposed multimodal approach to AED, a series of experiments have been conducted and their results presented in Table 2.

First, all sessions of isolated AEs were used to train the classifiers and the 3 scenes with spontaneously generated AEs were used for evaluation. The remaining 6 scenes were used for testing. In Table 2, the first column corresponds to baseline system with *one-against-all* topology. The next columns correspond to the results of feature level fusion with localization features, video features and combination of all modalities, respectively. As it can be observed, the video information improves the baseline results for those classes having a visual counterpart. This effect is justified by the fact that video information remains unaffected by acoustic noise. Therefore, the recognition rate of those classes considered as "difficult" (usually affected by overlap or of low energy) increases.

Acoustic localization features improve recognition accuracy for some AEs, but for other events, it is decreased. One of the reasons of such behavior is the mismatch between training and testing data. For instance, the cup clink AE in seminar conditions often appears when the person is standing, which is not the case for isolated AEs. Another reason is that, for overlapped AEs, the AE with higher energy will be localized while the other overlapped AE will be masked.

Finally, it has been observed that, for the paper work AE, the localization information reduces the accuracy rate, but when combined with video information, this rate increases. This effect is motivated by the complementarity of these two modalities. The opposite effect occurs with steps, where acoustic localization and video features together, decrease the overall performance.

6. Conclusions

In this work, by using data from interactive and rather spontaneous sessions, we have seen how video signals can be a useful additional source of information to cope with the problem of acoustic event detection. Acoustic localization features also tend to improve results for some particular classes. The combination of all these features for most of the classes produced higher recognition rates.

A *one-against-all* AED system architecture has been proposed, proving its effectiveness in the feature-level fusion approach. Future work will be devoted to extend the multimodal AED system to other classes as well as the elaboration of new multimodal features.

7. Acknowledgements

This work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first author is partially supported by a grant from the Catalan autonomous government. The first author would like to thank Anna Butko for her contribution to the paper.

8. References

[1] L. Lu, H. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation", IEEE Trans. on Speech and Audio Processing, vol. 10, pp. 504–516, 2002.

[2] T. Nishiura, S. Nakamura, K. Miki, and K. Shikano, "Environmental sound source identification based on hidden Markov models for robust speech recognition", in Proc. Eurospeech, pp. 2157–2160, 2003.

[3] A. Temko. "Acoustic event detection and classification", PhD thesis, Technical University of Catalonia, 2007.

[4] CHIL - Computers in the Human Interaction Loop – EU project. http://chil.server.de, 2004-2007.

[5] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, M. Omologo, "CLEAR Evaluation of Acoustic Event Detection and Classification systems", in Multimodal Technologies for Perception of Humans, LNCS, vol. 4122, Springer, 2007.

[6] T. Butko, A. Temko, C. Nadeu and C. Canton, "Fusion of Audio and Video Modalities for Detection of Acoustic Events", in Proc. Interspeech, pp. 123-126, 2008.

[7] C. Canton-Ferrer, T. Butko, C. Segura, X. Giró, C. Nadeu, J. Hernando, J.R. Casas, "Audiovisual Event Detection Towards Scene Understanding", in Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition, 2009.

[8] C. Nadeu, J. Hernando, and M. Gorricho, "On the decorrelation of filter-bank energies in speech recognition", in Proc. European Speech Processing Conference, pp. 1381–1384, 1995.

[9] R. Rifkin, A. Klautau, "In defense of One-Vs-All Classification", Journal of Machine learning Research, vol. 5, pp.101-141, 2004.

[10] J. DiBiase, H. Silverman, and M. Brandstein, „ Microphone Arrays: Techniques and Applications", M. S. Brandstein and D. B. Ward, Eds, pp. 157–180, Springer-Verlag, 2001.

[11] C. Canton-Ferrer, R. Sblendido, J. R. Casas, and M. Pardas, "Particle filtering and sparse sampling for multi-person 3D tracking", in Proc. IEEE Int. Conf. on Image Processing, pp. 2644–2647, 2008.

[12] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking", in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 252–259, 1999.

[13] X. Giró and F. Marqués, "Composite object detection in video sequences: Applications to controlled environments", in Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services, pp. 1–4, 2007.

[14] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bi-modal continuous speech recognition", in Proc. IEEE Workshop on Multimedia Signal Processing, pp. 65-70, 1998.