FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Detection of Overlapped Acoustic Events using Fusion of Audio and Video Modalities

Taras Butko and Climent Nadeu

TALP Research Center, Department of Signal Theory and Communications,
Universitat Politècnica de Catalunya, Barcelona, Spain
taras.butko@upc.edu, climent.nadeu@upc.edu

Abstract

Acoustic event detection (AED) may help to describe acoustic scenes, and also contribute to improve the robustness of speech technologies. Even if the number of considered events is not large, that detection becomes a difficult task in scenarios where the AEs are produced rather spontaneously and often overlap in time with speech. In this work, fusion of audio and video information at either feature or decision level is performed, and the results are compared for different levels of signal overlaps. The best improvement with respect to an audio-only baseline system was obtained using the feature-level fusion technique. Furthermore, a significant recognition rate improvement is observed where the AEs are overlapped with loud speech, mainly due to the fact that the video modality remains unaffected by the interfering sound.
Index Terms: Acoustic Event detection, Multimodal Fusion, Fuzzy Integral, Acoustic Localization

1. Introduction

Acoustic event detection (AED) aims at determining the identity of sounds and their temporal position in the signals that are captured by one or several microphones. It can provide a 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, speech enhancement.

AED is usually addressed from an audio perspective and many reported works are intended for indexing and retrieval of multimedia documents [1], or to improve robustness of speech recognition [2]. AED has been adopted as a relevant technology in several international projects, like CHIL [3], and evaluation campaigns [4]. The last international evaluations in seminar conditions have shown that AED is still a challenging problem. According to those results, the detection of AEs from only audio information shows a large amount of errors, which are mostly due to temporal overlaps of sounds.

The overlap problem may be faced by developing more efficient algorithms either at: the signal level, using source separation techniques like independent component analysis [5]; the feature level, by means of specific features [6]; or the model level [7]. An alternative approach consists of using an additional modality that is 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 rate. This idea was first presented in [8], 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 [9] [10] where the concept of multimodal AED is extended to detect and recognize a set of 11 AEs. In that work, not only video information but also acoustic source localization information was considered. Either a decision-level fuzzy integral fusion [9] or a feature-level fusion [10] was used to increase the accuracy of detection of isolated AEs. But for most of the AEs a statistically significant improvement was not observed due to the fact that in clean conditions the baseline recognition results are relatively high, so the additional modalities can not contribute significantly.

In this work we compare feature-level and decision-level fusion techniques for AED in more realistic conditions where the AEs are overlapped with speech. Feature-level fusion is performed by means concatenation of features from different modalities into one super-vector. Decision-level fusion is carried out with the Weighted Arithmetical Mean (WAM) approach and the Fuzzy Integral (FI) statistical approach [11].

2. Database and metric

There are several publicly available multimodal databases designed to recognize events, activities, and their relationships in interaction scenarios [3]. 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. In order to assess the performance of the proposed multimodal fusion approaches, the subset of isolated AEs from a recently recorded multimodal database [9] [10] was used. The video signals were recorded with 5 calibrated cameras at pixel resolution 768x576 and 25 fps. Audio signals were collected from 6 T-shaped 4-microphone clusters, and sampled at 44.1 kHz (in total, 24 microphones are used). All sensors were synchronized. In the recorded scenes, 5 different subjects performed several times the AEs employed in this work, adding up to around 100 instances for every AE, and 2 hours. This multimodal database is publicly available from the authors. We consider 12 classes of AEs which naturally occur in meeting-room environments, like in [7], [8], [9] and [10]: “Door knock”, “Door open/slam”, “Steps”, “Chair moving”, “Spoon/cup jingle”, “Paper work”, “Key jingle”, “Keyboard typing”, “Phone ring”, “Applause”, “Cough”, and “Speech”.

The meeting scenario adopted for this work assumes that there are two simultaneous acoustic sources in the room: one is always speech and the other is a specific AE. Taking into account this assumption, our UPC's smart-room has been considered ideally subdivided in the two areas: left and right (Figure 1 (a)). In the left part the speaker produces speech, and in the right part the listener produces different types of AEs. This assumption allows us to analyze the left and right parts of the room independently for the extraction of acoustic source localization features.

The speech of the speaker was recorded separately and it was artificially overlapped with the database of isolated AEs. To do that, for each AE instance, a segment with the same length was extracted from a random position inside the speech signal. The overlapping was performed with 5 different

Signal-to-Noise Ratios (SNRs): 20 dB, 10dB, 0dB, -10dB. -20 dB, where speech is considered as "noise".

Although the database with overlapped AEs is generated in an artificial way, it has some advantages:

a) The behavior of the system can be analyzed for different levels of overlap.

b) The existing databases of isolated AEs with high number of instances can be used for evaluations.

The metric referred to as AED-ACC [7], which is the F-score (harmonic mean between precision and recall), is employed to assess the final accuracy of the presented algorithms. .

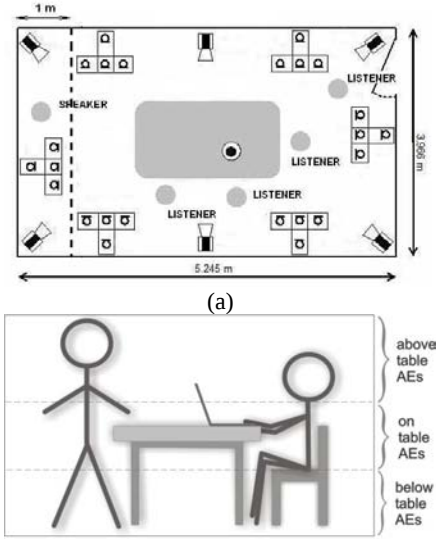

Figure 1: *(a) Top view of the room. (b) The three categories along the vertical axis.*

3. Feature extraction

A first stage of the proposed multimodal AED system is to determine the most informative features related to the AEs of interest for every input modality. Although audio and localization are originated from the same physical acoustic source, they are regarded as two different modalities.

3.1. Spectro-temporal audio features

A set of audio spectro-temporal features, like those used in automatic speech recognition, is extracted to describe every audio frame. It consists of 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives [10], which represent the spectral envelope of the audio waveform within a frame, as well as its temporal evolution. In total, a 32-dimensional feature vector is used. The FF feature extraction scheme consists in calculating a log filter-bank energy vector for each signal frame (in our experiments the frame length is 30 ms with 20 ms shift, Hamming window is applied) and then applying a FIR filter $h(k)$ on this vector along the frequency axis. We use the $h(k)=\{1, 0, -1\}$ filter in our approach. The end-points are taken into account. Notice that FF requires less computation than the classical MFCC.

3.2. 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 (Figure 1 (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 Figure 1 (b): *below table*, *on table*, *and above table.*

The acoustic localization system used in this work is based on the SRP-PHAT [12] 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.

3.3. 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 [13].

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 zenithal camera view, activity near the door can be addressed by means of a foreground/background pixel classification [14]. 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 zenithal 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 [15]. 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. Multimodal Acoustic Event Detection

Typically, low energy AEs such as paper wrapping, keyboard typing or footsteps are hard to detect using audio features, so both the visual correlate and the acoustic localization measures of these AEs may help to increase the detection performance.

In this paper, three data sources are combined for multimodal AED. First, two information sources are derived from acoustic data processing: single channel audio provides audio spectro-temporal (AST) features, while microphone array processing estimates the 3D location of the audio source. Second, information from multiple cameras covering the scenario allows extracting cues related to some AEs (described in Section 3.3). The features obtained from all modalities are combined together at feature and decision levels (Figure 2).

We employ a one-against-all detection strategy, so only two models are used for each AE, which will herewith be called "Class" and "non-Class". The first model is trained using the signals coming from the given class of interest,

while the second model is trained using the rest of signals. In total, 12 HMM-based binary detectors working in parallel are needed to perform detection of all AEs [10].

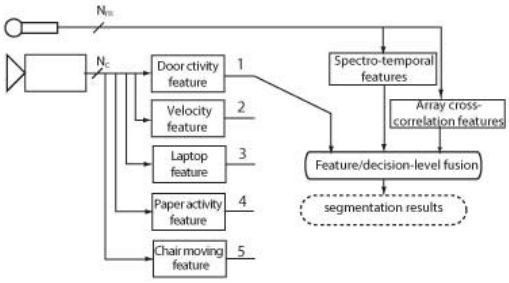

Figure 2: *System flowchart.*

4.1. Fusion of different modalities

The information fusion can be done on data, feature, and decision levels. Data fusion is rarely found in multi-modal systems because raw data is usually not compatible among modalities. For instance, audio is represented by one-dimensional vector of samples, whereas video is organized in two-dimensional frames. Concatenating feature vectors from different modalities into one super vector is an easy and simple way for combining audio and visual information. This approach has been reported, for instance, in [16] for multimodal speech recognition. An alternative to feature-level fusion is to model each different feature set separately, design a specialized classifier for this feature set, and combine the classifier output scores. Each such classifier acts as an independent "expert", giving its opinion about the unknown AE. The fusion rule then combines the individual experts' match scores. This approach is referred here as decision-level fusion. In the presented work, fusion is carried out on the decision level using weighted arithmetical mean (WAM) and fuzzy integral (FI) [11] fusion approaches. Unlike non-trainable fusion operators (mean, product), the statistical approaches WAM and FI avoid the assumption of equal importance of information sources. Moreover the FI fusion operator also takes into account the interdependences among modalities.

4.1.1. Feature-level fusion approach

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

$$\boldsymbol{Z} = \boldsymbol{X}_1 \cup \boldsymbol{X}_2 \cup ... \cup \boldsymbol{X}_S \qquad (1)$$

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

$$b_Z(t) = \sum_m p_m N(\boldsymbol{Z}_t; \boldsymbol{\mu}_m; \boldsymbol{\Sigma}_m). \qquad (2)$$

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

Feature-level fusion becomes a difficult task when some features are missing. Although the AST features can be extracted at every time instance, the feature that corresponds to the localization of acoustic source has an undefined value in the absence of any acoustic activity. In our experiments we substitute the missing features (*x*, *y*, *z* coordinates) with a predefined "synthetic" value (we use -1 value in our experiments). In this case we explicitly assign the 3D "position" of the silence event to have the value (*-1, -1, -1*).

4.1.2. Decision-level fusion approach

The decision-level fusion process is schematically depicted in Figure 3. First, a HMM segmentation based on the spectro-temporal features is performed to find all non-silence segments in the input audio. Given the "Class" and "non-Class" HMM models the log-likelihood ratio (LLR) is obtained for each non-silence segment S_i and each modality separately. A high positive LLR score would mean a high confidence that the non-Silence segment belongs to the "Class", while a low negative score would mean that the segment more likely belongs to "non-Class". A value close to zero indicates low confidence of decision. Furthermore, the obtained scores are normalized to be in the range [0…1] and their sum equal to 1. Then the normalized values are fused together using either Weighted Arithmetical Mean (WAM) or Fuzzy Integral (FI) fusion operators. To estimate the weights in WAM operator we use constrained regression approach to minimize the variance of error on development data [11]. The individual weights for the fuzzy integral fusion are also trained on development data using the gradient descent training algorithm.

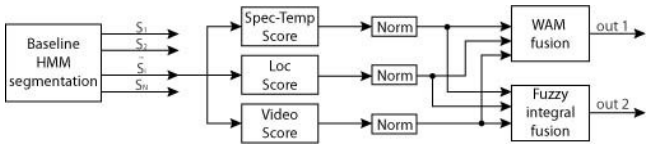

Figure 3: *Flowchart of decision-level fusion.*

5. Experiments and results

The detection results corresponding to two mono-modal AED systems based on AST and video features, respectively, are presented in Figure 4. The results for the video-based system are presented as an average accuracy score for those AEs for which the video counterpart is taken into consideration.

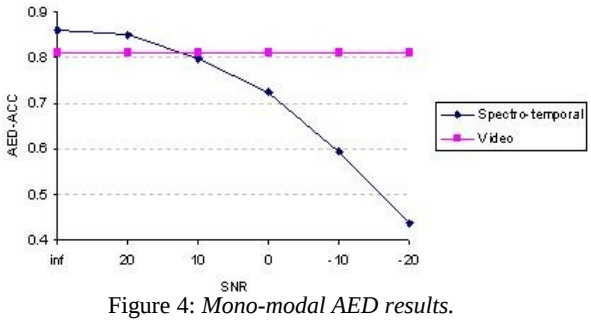

Figure 4: *Mono-modal AED results.*

Note the recognition results do not change for different SNR conditions since the video signals are not affected by overlapped speech. We do not present results for the AED system based on localization features since the information about the position of the acoustic source enables to detect just the category but not the AE within it. As we see from Figure 4, the recognition results of the baseline system decrease significantly for low SNRs.

The average relative improvement obtained by the multimodal system with respect to the baseline system (that uses the AST features only) for different fusion techniques is displayed in Figure 5. The feature-level fusion performs better for all AEs than both WAM and FI decision-level fusion approaches, and the both decision-level fusion techniques showed similar results in our experiments.

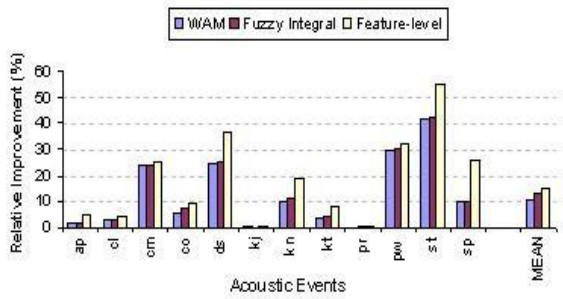

Figure 5: *Average relative improvement obtained by the multimodal system.*

The next Figure 6 summarizes (averaged over all Aes) the relative improvement obtained with the feature-level and decision-level (using fuzzy integral) fusion techniques for different levels of SNRs. According to these results, video signals as well as signals from arrays of microphones showed to be a useful additional source of information to cope with the problem of AED in overlapping conditions.

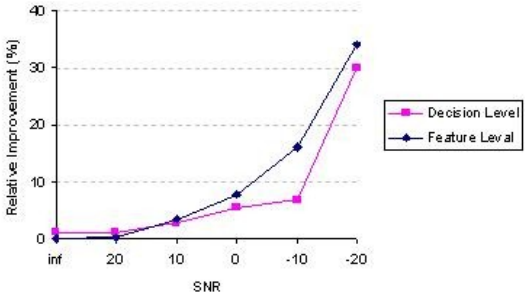

Figure 6: *The relative improvement obtained from multimodal features for different SNRs.*

6. Conclusions

In this paper, a comparison between multimodal systems based on a feature-level and decision-level fusion approaches have been presented. The acoustic data is processed to obtain a set of spectro-temporal features and the three localization coordinates of the sound source. Additionally, a number of features are extracted from the video signals by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of several AEs.

The obtained results showed that although in clean conditions the video and localization information does not contribute significantly, more improvement can be achieved in the conditions where the audio signals are overlapped with speech.

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. Particular thanks are given to Carlos Segura for providing the acoustic localization features in the proposed scenario. The authors would like to thank Cristian Canton-Ferrer and Xavier Giró for providing the features from the video signals.

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] CHIL: Computers in the Human Interaction Loop, <http://chil.server.de/>.

[4] CLEAR, 2007. Classification of Events, Activities and Relationships. Evaluation and Workshop. <http://www.clear-evaluation.org/>.

[5] A. Hyvärinen, J. Karhunen, E. Oja, “Independent Component Analysis”, John Wiley & Sons, 2001.

[6] S. Wrigley, G. Brown, V. Wan, S. Renals, “Speech and crosstalk detection in multi-channel audio”, IEEE Trans. Speech Audio Process, v. 13, pp. 84–91, 2005.

[7] A. Temko, C. Nadeu, “Acoustic event detection in meeting-room environments”, in Pattern Recognition Letters, v. 30, pp. 1281-1288, 2009.

[8] T. Butko, A. Temko, C. Nadeu and C. Canton, ”Inclusion of Video Information for Detection of Acoustic Events using the Fuzzy Integral”, in Machine Learning for Multimodal Interaction, LNCS, vol. 5237/2008, pp. 74-85, Springer, 2008

[9] 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.

[10] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, J.R. Casas, “Improving Detection of Acoustic Events Using Audiovisual Data and Feature Level Fusion”, in Proc. Interspeech, 2009.

[11] L. Kuncheva, Combining Pattern Classifiers, John Wiley & Sons, 2004.

[12] J. Dibiase, H. Silverman, M. Brandstein, “Microphone Arrays. Robust Localization in Reverberant Rooms”, Springer, 2001

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

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

[15] 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, 200.

[16] M. Chan, Y. Zhang, T. Huang, “Real-time lip tracking and bi-modal continuous speech recognition”, in Proc. IEEE Workshop on Multimedia Signal Processing, 1998.