Audiovisual Event Detection Towards Scene Understanding

C. Canton-Ferrer, T. Butko, C. Segura, X. Giró, C. Nadeu, J. Hernando, J.R. Casas
Technical University of Catalonia, Barcelona (Spain)
{ccanton,butko,csegura,xgiro,climent,javier,josep}@gps.tsc.upc.edu

Abstract

Acoustic events produced in meeting environments may contain useful information for perceptually aware interfaces and multimodal behavior analysis. In this paper, a system to detect and recognize these events from a multimodal perspective is presented combining information from multiple cameras and microphones. First, spectral and temporal features are extracted from a single audio channel and spatial localization is achieved by exploiting cross-correlation among microphone arrays. Second, several video cues obtained from multi-person tracking, motion analysis, face recognition, and object detection provide the visual counterpart of the acoustic events to be detected. A multimodal data fusion at score level is carried out using two approaches: weighted mean average and fuzzy integral. Finally, a multimodal database containing a rich variety of acoustic events has been recorded including manual annotations of the data. A set of metrics allow assessing the performance of the presented algorithms. This dataset is made publicly available for research purposes.

1. Introduction

Activity detection and recognition is a key functionality of perceptually aware interfaces working in collaborative human communication environments such as Smart Rooms. Within this context, the human activity is reflected in a rich variety of acoustic events (AEs), either produced by the human body or by objects handled by them, so auditory scene analysis [4] may help to detect and describe human activity, as well as producing informative outputs for higher semantic analysis systems.

Although speech is usually the most informative acoustic source, 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.

Acoustic Event Detection (AED) is usually addressed from an audio perspective and most of the existing contributions are intended for indexing and retrieval of multimedia documents [13] or to improve robustness of speech recognition [15]. Within the context of ambient intelligence, AED applied to give a contextual description of a meeting scenario was pioneered by [20]. Moreover, AED has been adopted as a semantically relevant technology in several international projects [1] and evaluation campaigns [21].

Overlapping between sounds is a typical problem faced by AED algorithms but can be tackled by employing additional modalities that are less sensitive to this phenomenon. Most of human produced AEs have a visual manifestation that can be exploited to enhance detection and recognition rates. This idea was first presented in [5] where the detection of footstep AE was improved by exploiting velocity information obtained from a visual person tracking system. In this paper, the concept of multimodal AED is extended to detect and recognize the set of AEs that commonly occur in a Smart Room scenario, namely applause, paper wrapping, footsteps, chair moving, coughing, door slaming, keyboard typing, door knocking, key jingling, phone ringing and cup clinking.

Three data sources are combined in this paper for multimodal AED. The overall operation of the proposed system is depicted in Fig.1. First, two information sources are derived from acoustic data processing: single channel audio provides spectral and temporal 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 involving several video-based technologies like person tracking, face detection, motion analysis, etc. The obtained features from all modalities are separately processed and a GMM-based classifier is trained for each of them. Finally, the outputs from these classifiers are combined using two decision level fusion techniques to be compared: the weighted mean average [12] and the fuzzy integral [5].

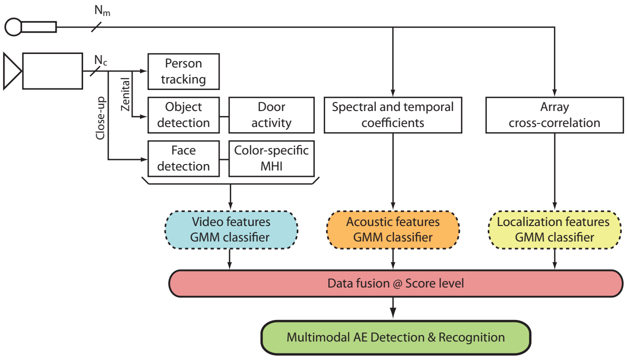

N_m
N_c
Zenital
Close-up
Person tracking
Object detection
Door activity
Face detection
Color-specific MHI
Spectral and temporal coefficients
Array cross-correlation
Video features GMM classifier
Acoustic features GMM classifier
Localization features GMM classifier
Data fusion @ Score level
Multimodal AE Detection & Recognition

Figure 1. System flowchart

A multi-camera and multi-microphone dataset containing a large number of instances of the AEs to be analyzed has been recorded and released for research purposes. Manual annotation of these data, together with some well accepted metrics [21], allowed testing the performance of the proposed algorithms proving the convenience of multimodal fusion in the AED task.

2. Monomodal Acoustic Event Detection

A first stage of our 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 in this paper. AEs presented in Tab.1 will be taken into account. The obtained features are afterward employed to train a classification module at each information source (audio, localization, and video).

2.1. Spectro-temporal audio features

When dealing with continuous audio streams, two approaches are found to analyze AEs [20]. The first one consists in detecting the AE endpoints by means of some heuristic rules and classifying the obtained audio segment afterward. The second approach classifies consecutive audio segments of fixed size producing a continuous output as the set of probabilities associated to every AE. Most AED systems prioritize this second technique due to its robustness and simplicity. Moreover, the detection task is converted into a classification problem.

This detection-by-classification technique becomes preferable when fusion is needed in future steps, since combining decisions made on audio segments of the same size is straightforward. According to Fig.2, some parameters have to be selected when applying this technique: the extracted features at every audio segment, the length of the analysis window, and the classification algorithm.

Acoustic Event	Audio	Localization	Video
Applause (ap)	+	+	+ (Hands motion)
Cup clink (cl)	+	+	
Chair moving (cm)	−	−	+ (Tracking)
Cough (co)	+	+	+ (Hands motion, Face detection)
Door slam (ds)	+	−	+ (Door activity)
Key jingle (kj)	+	+	
Door knock (kn)	+	+	
Keyboard typing (kt)	−	−	+ (Object detection)
Phone ringing (pr)	+	+	− (Hands motion)
Paper wrapping (pw)	−	−	+ (Paper motion)
Footsteps (st)	−	−	+ (Tracking)

Table 1. AEs analyzed in the present work together with their abbreviations. + and − express the detection complexity of every AE for every modality.

First of all, a set of spectro-temporal features are extracted to describe every audio frame. It consists of the 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives [14], which represent the spectral envelope of the audio waveform within the frame, as well as its temporal evolution within 5 consecutive frames. Regarding the analysis window, a Hamming window has been employed and its size empirically set to 30 ms. The window shift is set to 20 ms, that is, allowing some window overlap.

In automatic speech recognition, Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) classifiers are commonly employed. In both cases, audio segments are modelled via continuous density Gaussian mixtures. An alternative approach presented in [20] exploits Support Vector Machines (SVM) for binary classification. In the present work, we use a GMM-based classifier as it is well suited to model the audio segments of fixed length (unlike HMM) and it can be easily applied to a multi-class classification problem. Moreover, it explicitly provides the probabilities per each acoustic class for posterior multimodal fusion.

The obtained spectro-temporal features are used to train a GMM-based AE classifier using 5 Gaussians per each AE model with diagonal covariance matrices, using the expectation-maximization algorithm. Finally, the sequence of decisions is post-processed to get the detected events. In this step, the decisions are made on a 320 ms segment by assigning to the current decision segment the label that is most frequent.

2.2. Localization features

The spatial localization of a sound source can be used to enhance the detection of AEs. Although the global positions of the subjects in the analyzed scenario can be accurately retrieved using video information, this modality cannot easily determine whether a sound has been generated or at which z coordinate it has been produced, being this information a useful cue for AE classification.

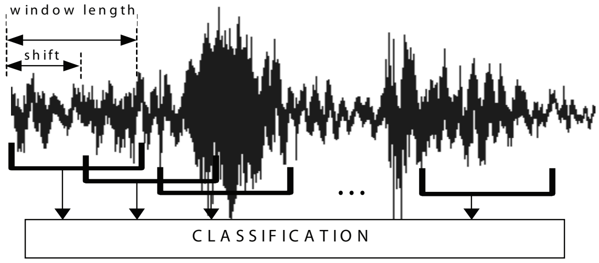


Figure 2. AE detection-by-classification approach.

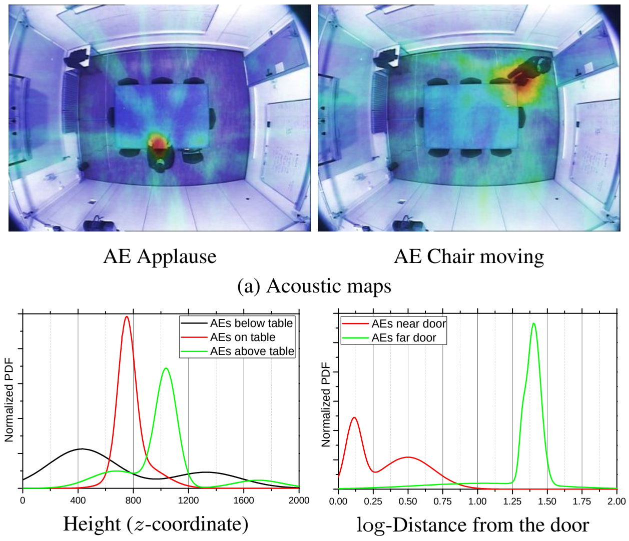
Many approaches to the task of acoustic source localization in smart environments have been proposed in the literature. Their main distinguishing characteristic is the way they gather spatial clues from the acoustic signals, and how this information is processed to obtain a reliable 3D position in the room space. Spatial features, like the Time Difference of Arrival (TDoA) between a pair of microphones [3] or the Direction of Arrival (DoA) of sound to a microphone array can be obtained on the basis of cross-correlation techniques [16], high resolution spectral estimation [17], or source-to-microphone impulse response estimation [8]. Depending on such features, the source position that agrees the most with the data streams and with the given geometry is selected. Conventional acoustic localization systems also include a tracking stage that smooths the raw position measurements to increase precision according to a motion model. However, these techniques require several synchronized high-quality microphones.

The acoustic localization system used in this work is based on the SRP-PHAT [9] 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 weighted cross-correlations from all the microphone pairs. The SRP-PHAT algorithm performs very robustly due to the PHAT weighting [16], and actually, it has turned out to be one of the most successful state-of-the-art approaches to microphone array sound localization.

Consider a scenario provided with a set of N_{M} microphones from which we choose a set microphone pairs, denoted as $\mathbb{S}$. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be the 3D location of two microphones i and j. The time delay of a hypothetical acoustic source placed at $\mathbf{x} \in \mathbb{R}^3$ is expressed as:

$$\tau_{\mathbf{x},i,j} = \frac{\|\mathbf{x} - \mathbf{x}_i\| - \|\mathbf{x} - \mathbf{x}_j\|}{s}, \tag{1}$$

where s is the speed of sound. The 3D space to be analyzed is quantized into a set of positions with typical separations of 5-10 cm. The theoretical TDoA $\tau_{\mathbf{x},i,j}$ from each exploration position to each microphone pair is precalculated and stored. PHAT-weighted cross-correlations of each microphone pair are estimated for each analysis frame [16]. They can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density ($G_{i,j}(f)$) as follows:



AE Applause

AE Chair moving

(a) Acoustic maps

(b) AE location distributions

Figure 3. Acoustic localization. In (a), acoustic maps corresponding to two AEs overlayed to a zenital camera view of analized scenario. In (b), the likelihood functions employed by the GMM classifier.

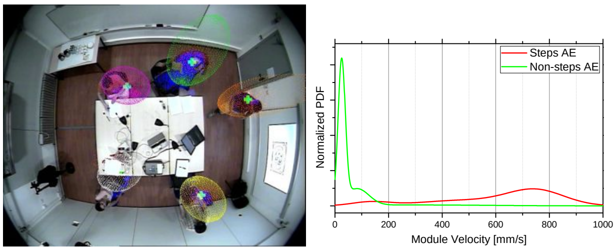
$$R_{i,j}(\tau) = \int_{-\infty}^{\infty} \frac{G_{i,j}(f)}{|G_{i,j}(f)|} e^{j2\pi f \tau} df. \tag{2}$$

The contribution of the cross-correlations is accumulated for each exploration region using the delays pre-computed in Eq.2. In this way, we obtain an *acoustic map* at every time instant, as depicted in Fig.3a. Finally, the estimated location of the acoustic source is the position of the quantized space that maximizes the contribution of the cross-correlation of all microphone pairs:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \sum_{i,j \in \mathbb{S}} R_{i,j}(\tau_{\mathbf{x},i,j}). \tag{3}$$

The sum of the contributions of each microphone pair cross-correlation gives a value of confidence of the estimated position, which is assumed to be well-correlated with the likelihood of the estimation given.

Localization of a given sound does not allow distinguishing among AEs that are produced around the same height (z coordinate), such as between cup clinking and phone ringing or between footsteps and chair moving. Indeed, every AE has an associated height, allowing the following AE taxonomy: *below table*, *on table* and *above table*. In the xy plane, only those AEs physically related with the environment can exploit this information and, in our case, we selected the door as a reference, defining two AE meta-classes: *near door* and *far door*. This information is employed to train a GMM-based classifier using 3 Gaussians per class as depicted in Fig.3b.



(a)

(b)

Figure 4. Person tracking. In (a), the output of the employed algorithm in a scenario involving multiple targets. In (b), the likelihood functions used by the GMM classifer.

2.3. Video features

Information gathered from multiple views can be processed towards detecting certain actions associated with some AEs. In this section, several technologies providing useful cues for AED are presented.

Person tracking

Tracking of multiple people present in the analysis area basically produces two figures associated with each target: position and velocity. As it has been commented previously, acoustic localization is directly associated with some AEs but, for the target's position obtained from video, this assumption cannot be made. Nonetheless, target's velocity is straightforward associated with footstep AE. Once the position of the target is known, an additional feature associated with the person can be extracted: height. When analyzing the temporal evolution of this feature, sudden changes of it are usually correlated with chair moving AE, that is, when the person sits down or stands up.

Multiple cameras are employed to perform tracking of multiple interacting people in the scene, applying the real-time performance algorithm presented in [6]. This technique exploits spatial redundancy among camera views towards avoiding occlusion and perspective issues by means of a 3D reconstruction of the scene. Afterward, an efficient Monte Carlo based tracking strategy retrieves an accurate estimation of both the location and velocity of each target at every time instant. An example of the performance of this algorithm is shown in Fig.4a. A GMM-based classifier is employed to detect steps from the obtained velocity as shown in Fig.4b.

Color-specific MHI

Some AEs are associated with motion of hands or objects around the person. In particular, we would like to detect hands motion in the horizontal direction that can be correlated with applause AE. Assuming a “polite” environment, where people place the hand in front of the mouth before coughing, vertical hand motion can be associated with this AE. Along the same lines, a motion of a white object in the scene can be associated to paper wrapping (under the assumption that a paper sheet is distinguishable from the background color). In order to address the detection of these motions, a close-up camera focused on the front the person under study is employed.

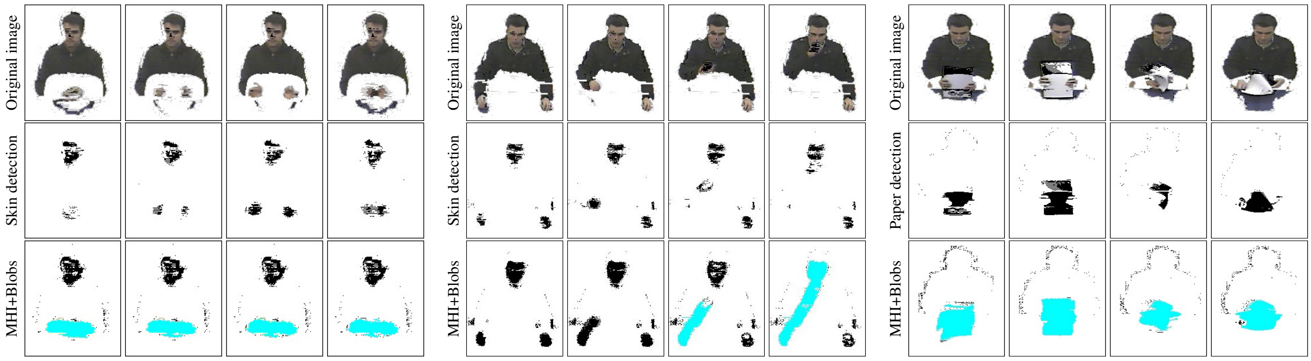
Motion descriptors introduced by [2], namely the motion history energy (MHE) and image (MHI), have been found useful to describe and recognize actions. In this paper, only the MHE feature is employed where every pixel in the MHE image contains a binary value denoting whether motion has occurred in the last τ frames at that location. In the original technique, silhouettes were employed as the input to generate these descriptors but not being appropriate in our context since motion typically occurs within the silhouette of the person. Instead, we propose to generate the MHE from the outputs of a pixel-wise color detector, hence performing a color/region-specific motion analysis. For the hands motion analysis, the histogram-based skin color detector proposed in [11] is employed. For paper motion, a statistic classifier based on a Gaussian model in RGB is used to select the pixels with whitish color. In our experiments, $\tau = 12$ frames produced satisfactory results.

Finally, a connected component analysis is applied to the obtained MHE images and some properties are computed over the retrieved components (blobs). The area of each blob allows discarding spurious motion. In the hand motion case, the major and minor axis relation of the ellipse fitted to the biggest blob in the scene, together with the orientation of this blob, allows distinguishing horizontal motion (applause AE) and vertical motion (cough or phone ringing AE). In the paper motion case, the size of the biggest blob in the scene is employed to address paper wrapping AE detection. An example of this technique is depicted in Fig.5. As in the previous cases, a GMM-based system is employed to classify the motion produced in this close-up view of the subject under study.

Object detection

Detection of certain objects in the scene can be beneficial to detect some 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.

Captured images are segmented to create an initial partition of 256 regions based on color similarity. These regions are iteratively fused to generate a Binary Partition



(a) Clapping (b) Coughing (c) Paper wrapping

Figure 5. Color-specific motion analysis. In (a) and (b), two examples of skin color based horizontal and vertical motion detection, associated to two AEs. In (c), white color motion detection is associated with paper wrapping.

Tree (BPT), a region-based representation of the image that provides segmentation at multiple scales [18]. Starting from the initial partition, the BPT is built by iteratively merging the two most similar and neighbouring regions, defining a tree structure whose leaves represent the regions at the initial partition and the root corresponds to the whole image (see Fig.6b). Thanks to this technique, the laptop parts may be detected not only at the regions in the initial partition but also at some combinations of them, represented by the BPT nodes. Once the BPT is built, visual descriptors are computed for each region represented at its nodes. These descriptors represent color, area and location features of each segment.

The detection problem is posed as a traditional pattern recognition case, where a GMM-based classifier is trained for the screen and keyboard parts. A subset of ten images representing the laptop at different positions in the table has been used to train a model based on the region-based descriptors of each laptop part, as well as their relative position and sizes. An example of the performance of this algorithm is shown in Fig.6a. For further details on the algorithm, the reader is referred to [10].

Face detection

Face detection on the close-up view of the person under study has been considered using the standard algorithm described in [22]. Face position has been found relevant in cough AE, since vertical hand motion usually ends at the face in the act of covering the mouth while coughing. This position has been employed to define a new feature, based on the distance from the center of the face to the end of the vertical hand motion thus providing an extra feature for visual cough AE detection. Phone ringing AE also be addressed using the same feature since usually the user places the phone near the face when talking.

Door activity

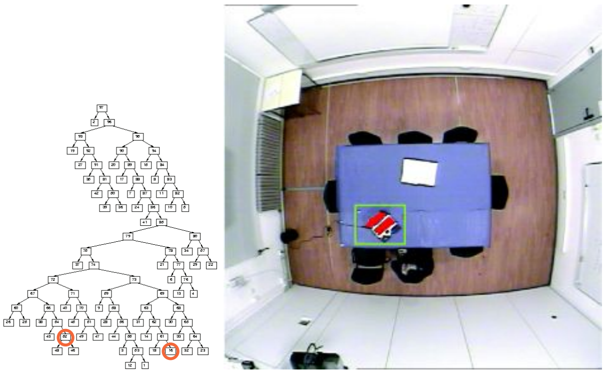
In order to visually detect door slam AE, we considered exploiting the *a priori* knowledge about the physical location of the door. Analyzing the zenital camera view, activity near the door can be addressed by means of a foreground/background pixel classification [19]. The amount of foreground pixels in the door area will indicate that a person has entered or exited hence allowing a visual detection of door slam or door knock AE.

3. Multimodal Acoustic Event Detection

Typically, low acoustic energy AEs such as paper wrapping, keyboard typping or footsteps are hard to be detected using audio features while the visual occurrence of these AEs is well correlated with the output of some video processing algorithms thus justifying our multimodal approach to AED. However, some considerations must be taken into account.

3.1. Fusion of different modalities

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 in [7] for multimodal speech recognition. Although feature-level fusion may improve recognition accuracy, it has several shortcomings. First, fusion becomes difficult if a feature is missing (e.g. velocity of the person in the room while nobody is inside). Second, the number of training vectors needed for robust density estimation increases exponentially with the dimensionality.



(a) (b)

Figure 6. Object detection. In (a) the binary partition tree representing the whole image as a hierarchy. Regions corresponding to the screen and keyboard regions are identified within the tree. In (b), the detection of a laptop from the zenital view.

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 of the different feature sets 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 presented work, fusion is carried out on decision-level using weighted arithmetical mean (WAM) [12] and fuzzy integral (FI) [20] fusion approaches. Unlike non-trainable fusion operators (mean, product), the statistical WAM and FI approaches overcome assumption about equal importance of information sources. Moreover FI fusion operator also takes into account interdependences among modalities.

3.2. Synchronization and normalization

In order to fuse 3 information sources their outputs must be synchronized in time. In our case, the detection system based on spectro-temporal features provides decisions every 20 ms, while the outputs of detection systems based on localization and video features are every 40 ms. To overcome the problem of misalignment, the decisions of each system were interpolated to a common time step of 20 ms. For that purpose, the outputs of the video-based and localization-based systems were replicated twice.

While the spectro-temporal AED system provides probabilities for each AE, acoustic localization and video-based systems provide probabilities for group of classes (meta-classes) such as "below table", "on table", "vertical motion", etc. To make fusion possible, it is necessary to distribute (extend) the meta-class score to all the classes inside the meta-class. We do it by means of assigning the same score to all the classes inside the meta-class. In the case when AE belongs to n different meta-classes, the final score is computed as the product of the n individual scores corresponding to each meta-class.

In order to make the outputs of information sources commensurable, we have to normalize them to be in the range $[0, 1]$ and their sum equal to 1. The soft-max function is then applied to the vector of scores of each detection system. This function is defined as:

$$\hat{q}_i = \frac{\exp(k \cdot q_i)}{\sum_{j=1}^{M} \exp(k \cdot q_j)}, \tag{4}$$

where the coefficient k controls the distance between the components of the vector $[q_1, q_2, \ldots, q_M]$. For instance, in extreme case when $k = 0$, the elements of the vector after soft-max normalization would have the same value M^{-1}, and when $k \to \infty$ the elements tend to become binary.

3.3. WAM and FI fusion schemes

We are searching for a suitable fusion operator to combine a finite set of information sources $Z = \{1, \ldots, z\}$. Let $D = \{D_1, D_2, \ldots, D_z\}$ be a set of trained classification systems and $\Omega = \{c_1, c_2, \ldots, c_N\}$ be a set of class labels. Each classification system takes as input a data point $x \in \mathbb{R}^n$ and assigns it to a class label from Ω. Alternatively, each classifier output can be formed as an N-dimensional vector that represents the degree of support of a classification system to each of N classes. We suppose these classifier outputs are commensurable, i.e. defined on the same measurement scale (most often they are posterior probability-like).

Let us denote h_i, $i = 1, \ldots, z$, as the output scores of z classification systems for the class c_n. The WAM fusion operator is defined as:

$$M_{\text{WAM}} = \sum_{i \in Z} \mu(i) h_i, \tag{5}$$

where $\sum_{i \in Z} \mu(i) = 1$ and $\mu(i) \geq 0$, $\forall i \in Z$. The WAM operator combines the score of z competent information sources through the importance weights $\mu(i)$. However, the main disadvantage of the WAM operator is that it implies preferential independence of the information sources.

Assuming the sequence h_i is ordered in such a way that $h_1 \leq \cdots \leq h_z$, the Choquet *fuzzy integral* can be computed as:

$$M_{\text{FI}} = \sum_{i=1}^{z} [\mu(i, \ldots, z) - \mu(i+1, \ldots, z)] h_i, \tag{6}$$

where $\mu(z+1) = \mu(\emptyset) = 0$. $\mu(S)$ can be viewed as a weight related to a subset S of the set Z of information sources. It is called *fuzzy measure* (FM) and for $S, T \subseteq Z$ has to meet the following conditions:

$$\mu(\emptyset) = 0, \mu(Z) = 1, \qquad \text{(Boundary)} \tag{7}$$

$$S \subseteq T \Rightarrow \mu(S) \leq \mu(T). \qquad \text{(Monotonicity)} \tag{8}$$

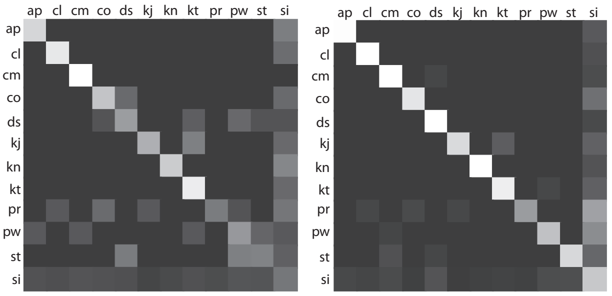
The large flexibility of the FI aggregation operator is due to the use of the FM that can model importance of criteria. Although the FM $\mu(i)$ provides an initial view about the importance of information source i, all possible subsets of Z that include that information source should be analyzed to give a final score. For instance, we may have $\mu(i) = 0$, suggesting that element i, $i \neq T$, is not important; but if, at the same time, $\mu(T \cup i) \gg \mu(T)$, this actually indicates that i is an important element for the decision. To calculate the importance of the information source i, the Shapley [20] score is employed.

4. Dataset and Experiments

In order to assess the performance of the proposed multimodal AED system, a dataset has been recorded with 5 calibrated cameras at a resolution of 768x576 at 25 fps. 6 T-shaped 4-microphone clusters are also employed sampling at 44kHz. Synchronization among all sensor is fulfilled. In the recorded scenes, 4 subjects performed the 11 AEs employed in this work several times, adding up to 50 instances for every AE. The recorded dataset has been divided into two parts for training and testing adding up to 2 hours of data of events. Manual annotation of the data has been done towards a fair performance evaluation. In order to encourage other researchers into the multimodal AED field, this dataset is made publicly available[1].

The metric defined in [21] 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. These figures are defined as follows: *precision* is the number of correct hypotheses AEs divided by the total number of hypotheses AEs and *recall* is the number of correctly detected reference AEs divided by the total number of reference AEs.

For all the AEs presented in this paper, a series of experiments have been conducted towards detecting them using the three presented information sources: audio, localization and video. The obtained results are presented in Tab.2. In the same table, importance of each information source, computed through the Shapley measure, is also presented for every AE. The detection of some low energy AEs has improved with reference to the baseline when adding different modalities. In the case of footsteps, there is a relative improvement of 244%, basically due to the video contribution (as pointed out by the importance of this source). Paper wrapping has also benefited from multimodality with a 15% relative detection performance increment. Other AEs have slightly improved their detection rates thus finally achieving an AEs average relative improvement of 7.5%. Confusion matrices obtained for the acoustic baseline AED and for the FI AED combining the three information sources is shown in Fig.7, where multimodality has contributed to reduce the inter-class confusion.



(a) Baseline AED (b) A+L+V FI-AED

Figure 7. Confusion matrices. Silence AE (si) is displayed although not being explicitly addressed in this paper.

When comparing the contribution of every information source to the overall performance, it can be noticed that only for certain AEs (keyboard typing, paper wrapping or footsteps), location and video sources contribute to improve the detection rate. Results obtained with WAM or FI are different in this case thus pointing out the adequateness of FI to fuse multimodal data. In the rest, the detection rate does not experience a significant change.

5. Conclusions and Future Work

This paper presents a novel multimodal approach to acoustic event detection relying not only on audio information but also on localization and video data. Acoustic information is processed to obtain a set of spectro-temporal features and a 3D localization of the sound source. From the video side, a number of systems aim at detecting the visual counterpart of AEs by means of object and face detection, motion analysis, or multi-camera person tracking. Two schemes are proposed to fuse these three information inputs, namely the weighted mean average and the fuzzy integral. Finally, experiments conducted over an annotated database proved that for some low energy AEs (paper wrapping, footsteps, and keyboard typing), multimodality yields a noticeable increase of the detection rate.

Systems presented in this paper, especially the video ones, do not output a feature at every time instant. This effect produces variable size feature vectors turning out multimodal data fusion techniques at feature level difficult and therefore becoming our inmediate future research direction. Using asynchronous HMM based classfiers are also under study. Other future work lines aim at analyzing the psychological aspects of AED/C and its implications in focus of attention estimation.

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

		WAM			FI			Source importance		
Acoustic Event	Audio Baseline	A+V	A+L	A+L+V	A+V	A+L	A+L+V	A	L	V
Applause	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.46	0.31	0.22
Cup clink	0.87	0.87	0.87	0.87	0.87	0.87	0.91	0.56	0.24	0.19
Chair moving	0.89	0.89	0.89	0.89	0.89	0.89	0.87	0.48	0.20	0.30
Cough	0.78	0.78	0.79	0.78	0.78	0.79	0.79	0.46	0.31	0.21
Door slam	0.92	0.96	0.91	0.95	0.89	0.91	0.84	0.22	0.25	0.51
Key jingle	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.50	0.26	0.23
Door knock	0.91	0.91	0.92	0.92	0.91	0.92	0.92	0.50	0.25	0.23
Keyboard typing	0.84	0.86	0.86	0.88	0.85	0.86	0.91	0.48	0.21	0.29
Phone ringing	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.44	0.29	0.26
Paper wrapping	0.74	0.81	0.73	0.82	0.81	0.84	0.85	0.31	0.25	0.42
Footsteps	0.18	0.64	0.20	0.65	0.64	0.22	0.81	0.23	0.13	0.62
Average	0.80	0.85	0.80	0.85	0.85	0.81	0.86	0.42	0.24	0.34

Table 2. Multimodal AED results for the WAM and FI fusion methods using audio, localization and video. Importance of every information source is also displayed.

References

[1] CHIL - Computers in the Human Interaction Loop - EU project. http://chil.server.de, 2004-2007. 1

[2] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23:257–267, 1999. 4

[3] M. Brandstein. *A Framework for Speech Source Localization Using Sensor Arrays*. PhD thesis, Brown University, 1995. 3

[4] A. Bregman. *Auditory Scene Analysis*. MIT Press, 1990. 1

[5] T. Butko, A. Temko, C. Nadeu, and C. Canton-Ferrer. Fusion of audio and video modalities for detection of acoustic events. In *Proc. Interspeech*, 2008. 1

[6] C. Canton-Ferrer, R. Sblendido, J. R. Casas, and M. Pardàs. Particle filtering and sparse sampling for multi-person 3D tracking. In *Proc.IEEE Int. Conf. on Image Processing*, pages 2644–2647, 2008. 4

[7] M. Chan, Y. Zhang, and T. Huang. Real-time lip tracking and bi-modal continuous speech recognition. In *Proc. IEEE Workshop on Multimedia Signal Processing*, 1998. 5

[8] J. Chen, N. Huang, and J. Benesty. An adaptive blind SIMO identification approach to joint multichannel time delay estimation. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 4, pages 53–56, 2004. 3

[9] J. DiBiase, H. Silverman, and M. Brandstein. *Microphone Arrays. Robust Localization in Reverberant Rooms*. Springer, 2001. 3

[10] 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*, pages 1–4, 2007. 5

[11] M. Jones and J. Rehg. Statistical color models with application to skin detection. *Int. Journal of Computer Vision*, 46:1:81–96, 2002. 4

[12] L. Kuncheva. *Combining Pattern Classifiers*. John Wiley & Sons, 2004. 1, 6

[13] L. Lu, H. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Trans. on Speech and Audio Processing*, 10:7:504–516, 2002. 1

[14] C. Nadeu, J. Hernando, and M. Gorricho. On the decorrelation of filter-bank energies in speech recognition. In *Proc. European Speech Processing Conference*, pages 1381–1384, 1995. 2

[15] 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*, pages 2157–2160, 2003. 1

[16] M. Omologo and P. Svaizer. Use of the crosspower-spectrum phase in acoustic event location. *IEEE Trans. on Speech and Audio Processing*, 5:3:288–292, 1997. 3

[17] I. Potamitis, G. Tremoulis, and N. Fakotakis. Multi-speaker DOA tracking using interactive multiple models and probabilistic data association. In *Proc. European Conference on Speech Communication and Technology*, 2003. 3

[18] P. Salembier and L. Garrido. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE Trans. on Image Processing*, 9:4:561–576, 2000. 5

[19] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 252–259, 1999. 5

[20] A. Temko. *Acoustic event detection and classification*. PhD thesis, Technical University of Catalonia, 2007. 1, 2, 6, 7

[21] A. Temko, C. Nadeu, and J. Biel. Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07. In *Proceedings of Classification of Events, Activities and Relationships Evaluation and Workshop*, volume 4625 of *Lecture Notes on Computer Science*, pages 354–363, 2007. 1, 2, 7

[22] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001. 5