TWO-SOURCE ACOUSTIC EVENT DETECTION AND LOCALIZATION: ONLINE IMPLEMENTATION IN A SMART-ROOM

Taras Butko, Fran González Pla, Carlos Segura, Climent Nadeu, and Javier Hernando

Department of Signal Theory and Communications, TALP Research Center,
Universitat Politècnica de Catalunya
Campus Nord, c/ Jordi Girona, 1-3, Barcelona, Spain
{taras.butko, frangp, csegura, climent.nadeu, javier.hernando}@upc.edu

ABSTRACT

Real-time processing is a requirement for many practical signal processing applications. In this work we implemented online 2-source acoustic event detection and localization algorithms in a Smart-room, a closed space equipped with multiple microphones. Acoustic event detection is based on HMMs that enable to process the input audio signal with very low latency; acoustic source localization is based on the SRP-PHAT localization method which is known to perform robustly in most scenarios. The experimental results from online tests show high recognition accuracy for most of acoustic events both isolated and overlapped with speech.

1. INTRODUCTION

Activity detection and description is a key functionality of perceptually aware interfaces working in collaborative human communication environments like meeting-rooms or classrooms. Actually, in the context of person-machine communication, computers involved in human communication activities have to meet certain requirements and be designed to have minimal possible awareness from the users. Consequently, there is a need of perceptual user interfaces which are multimodal and robust, and which use unobtrusive sensors that should sense the ongoing human activity. As human activity is reflected in a rich variety of acoustic events, either produced by the human body or by objects handled by humans, acoustic event detection (AED) may help to describe the human and social activity. Ringing telephones, clapping or laughter inside a speech discourse, a strong yawn in the middle of a lecture, knocks on doors, doors opening and closing, footsteps, or even the difference between one person speaking or more people speaking at the same time, are auditory cues that can be used to detect relevant events and state changes on meetings.

For meeting-room environments, the task of AED is relatively new; however, it was already evaluated in the framework of two international evaluation campaigns: in CLEAR (Classification of Events, Activities, and Relationships evaluation campaigns) 2006 [1], by three participants, and in CLEAR 2007 [2], by six participants. In the last evaluations, 5 out of 6 submitted systems showed accuracies below 25%, and the best system got 33.6% accuracy [3]. In most submitted systems, the standard combination of cepstral coefficients and hidden Markov model (HMM) classifiers, widely used in speech recognition, was exploited. It was found that the overlapping segments account for more than 70% of errors produced by every submitted system. It was clear since then that the detection of overlapped AEs was a challenging task in the context of meeting-room AED [4][5].

In the work reported here we implemented both an HMM-based AED system and an acoustic source localization (ASL) system operating in real-time (on-line) in the UPC's smart-room using the signals captured by the set of distant microphones available in that room. The proposed system is able to detect not only isolated AEs but also AEs overlapped with speech. The problem of signal overlaps is dealt with at the level of models [6]: additional acoustic models for signal overlaps are considered for both training and testing.

There was a first online implementation of AED and ASL technologies in our smart-room in the context of the European CHIL project [7] in 2007. That one-source AED was implemented using support vector machines (SVMs) [6]. An improved version was developed later, in 2009 [8], where the system was able to detect not only isolated AEs but also AEs overlapped with speech. In those SVM realizations, the AED was performed by sequential classification of 1sec sliding windows with 0.2 sec shift. In the work reported here we propose an alternative algorithm for AED based on HMMs, where the acoustic analysis is performed on a frame-by-frame basis, using the Viterbi segmentation algorithm for recognition. Besides, the acoustic source localization algorithm is extended to 2 sources.

2. SCENARIO, DATABASES AND EXPERIMENTAL SETUP

In our work we consider 12 classes of AEs which naturally occur in meeting-room environments, like in [4], [5], [6] and [8] (Table 1).

The UPC's smart-room (Figure 1) is a closed space equipped with multiple microphones and cameras, which was designed to assist human activities. It provides infrastructure for doing research on audio-visual perception technologies.

The meeting scenario adopted for this work assumes 3 different modes:

- there is no acoustic activity
- there is only one acoustic source in the room
- there are two simultaneous acoustic sources, one of which is always speech

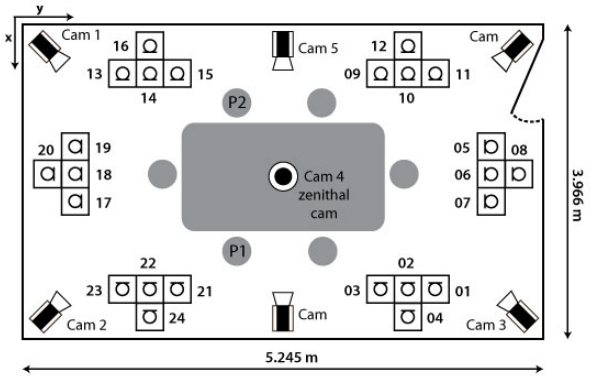


Figure 1 – The UPC smart-room.

Table 1 – Number of occurrences per acoustic event class for the training and testing data.

Event Type	label	Number of Occurrences
Door knock	[kn]	79
Door open/slam	[ds]	256
Steps	[st]	206
Chair moving	[cm]	245
Spoon/cup jingle	[cl]	96
Paper work	[pw]	91
Key jingle	[kj]	82
Keyboard typing	[kt]	89
Phone ring	[pr]	101
Applause	[ap]	83
Cough	[co]	90
Speech	[sp]	74

In the case when there are two simultaneous acoustic sources, we assume that speech is always at the right part of the room (a speaker close to the blackboard), and the other AE is at the left part (people placed around the table). This assumption enables to associate each of the two acoustic sources with the coordinates provided by the ASL system.

The database used to train and test the models consists of the audio part of the publicly available multimodal database used in [5]. The number of acoustic event instances for each isolated AE is displayed in Table 1. The database of AEs overlapped with speech was artificially generated using speech recorded separately. 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 3 different signal-to-noise ratios (SNRs): 10dB, 0dB, -10dB, where speech is considered as "noise".

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

- The behavior of the system can be analyzed for different levels of SNR.
- The existing databases of isolated AEs with high number of instances can be used for training and testing.

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

3. METHODOLOGY

The flowchart of the proposed online AED-ASL system is depicted in Figure 2. It consists of 4 main blocks: audio acquisition, 2-source AED, 2-source ASL and visualization.

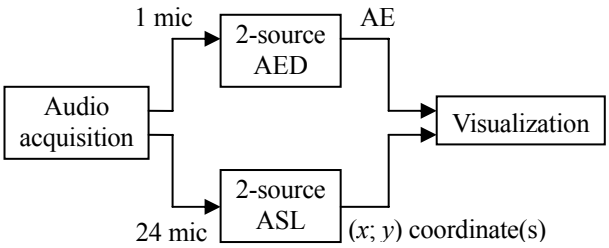


Figure 2 – The flowchart of the proposed AED-ASL system working online.

A software package called SmartAudio++ developed under Linux platform was used to implement those components. In the audio acquisition block, the audio signals are captured simultaneously from 24 microphones from T-shape clusters located on the walls of the room. The audio signal from the microphone #18 is used for subsequent feature extraction and recognition in the 2-source AED block; the 2-source ASL block uses the whole set of 24 microphones to estimate the position(s) of the acoustic source(s). The microphone #18 is used because it is the nearest one to the table. The output from the 2-source AED block is either an isolated AE (if one source is detected) or an AE overlapped with speech (if 2 sources are detected). The output from the 2-source ASL block is either one or two sets (*x, y*) of coordinates of the acoustic source(s). Both outputs are combined together and visualized by a graphical user interface (GUI). An ambiguity happens when the number of acoustic sources detected by the AED and ASL blocks is different. In this case the number of the displayed acoustic sources corresponds to the number of acoustic sources detected by the AED block. The AEs are displayed in default positions if the number of localization coordinates is less than the number of detected AEs.

3.1 Two-source AED system

The first step in our 2-source AED system is feature extraction. A set of audio spectro-temporal features is extracted to describe every audio signal frame. In our experiments, the frame length is 30 ms with 20 ms shift, and a Hamming window is applied. There exist several alternative ways of parametrically representing the spectrum envelope of audio signals. The mel-cepstrum representation is the most widely used in recognition tasks. In our work, we employ a variant of them called frequency-filtered (FF) log filter-bank ener-

gies (LFBE) [9]. It consists of applying, for every frame, a short-length FIR filter to the vector of log filter-bank energies vector, along the frequency variable. The transfer function of the filter is $z\text{-}z^{-1}$, and the end-points are taken into account. That type of features have been successfully applied not only to speech recognition but also to other speech technologies like speaker recognition [10]. In the experiments, 16 FF-LFBEs are used, along with their first temporal derivatives, the latter representing the temporal evolution of the envelope. Therefore, a 32-dimensional feature vector is used.

We use a hidden Markov model (HMM) based AED system like the ones used for continuous speech recognition, where Gaussian mixture models (GMM) are used to compute the state emission probability [11][5]. The HTK toolkit [12] is used for training and testing the HMM–GMM system. There is one HMM for each AE, with only one emitting state, a topology that showed the best results using a cross-validation procedure on the development data. Hence, GMMs are actually used, and the HMM formalism is only concerned with the actual implementation. 64 Gaussian components with diagonal covariance matrix are used per model. Each HMM is trained with the signal segments belonging to the corresponding event class using the standard Baum–Welch training algorithm [11]. In total, 24 HMMs are trained, one for each isolated AE class and one for each AE class overlapped with speech. For testing, the Viterbi algorithm is used. In the online system implementation we use ATK, an API designed to facilitate building experimental applications with HTK.

In the proposed implementation, the AED recogniser always operates in one of three possible states as indicated by the state diagram shown in Figure 3. Initially, when there is no acoustic activity in the room, the recognizer operates in FLUSH state and discards the input audio packets (frames). When the energy of the N consecutive packets exceeds the predefined threshold, the recognizer starts operating in RUN state. Similarly, the recognizer goes back to FLUSH state if N consecutive packets are below this threshold. In RUN state the recognizer continuously performs the Viterbi decoding of the waveform within the time interval $[t_0\ t_i]$, where t_0 is the time instance when the first non-silence packet is received, and t_i is the current time stamp. The defined grammar, illustrated in Figure 4 allows only one AE to be detected on that interval. In the ANS state the recognizer sends the current decision obtained in the RUN state to the visualization block. There are 2 possible conditions for the recognizer going to the ANS state:

- The confidence of the current decision exceeds a predefined threshold. In this case the corresponding AE label is sent to the visualization block. Very often a confident decision is obtained with just a few input packets. In this case the time delay between the AE production and its visualization is small.
- During 1 second of operation in the RUN state, a decision with enough confidence is not obtained. In this case the "unknown" AE is sent to the visualization block.

The recognizer goes back to the RUN state immediately when the output label is sent to the visualization block.

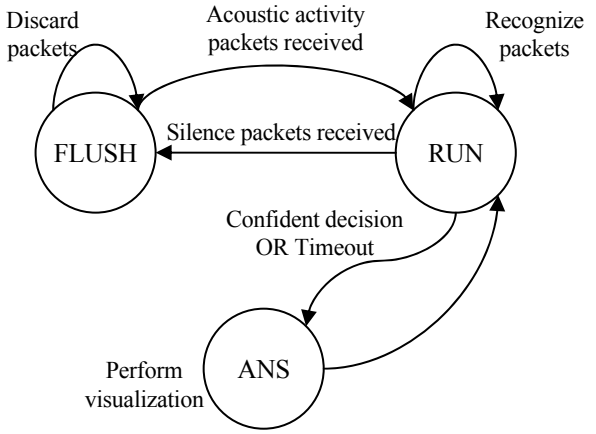


Figure 3 – Finite-state machine.

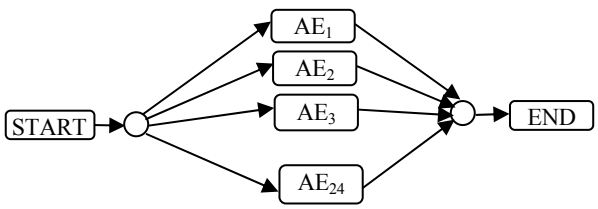


Figure 4 – AED grammar.

3.2 Two-source acoustic source localization system

The acoustic localization system used in this work is based on the SRP-PHAT [13] localization method, which is known to perform robustly in most scenarios. The SRP-PHAT algorithm is briefly described in the following. Consider a scenario provided with a set of N_M microphones from which we choose a set microphone pairs, denoted as Ψ. Let X_i and X_j be the 3D location of two microphones i and j. The time delay of a hypothetical acoustic source placed at $x \in R^3$ is expressed as:

$$\tau_{x,i,j} = \frac{\left\|x - x_i\right\| - \left\|x - x_j\right\|}{s} \qquad (1)$$

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 to 10 cm. The theoretical TDoA $\tau_{x,i,j}$ from each exploration position to each microphone pair is pre-calculated and stored. PHAT-weighted cross-correlations of each microphone pair are estimated for each analysis frame [14]. They can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density $G_{i,j}(f)$ as follows:

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

The contribution of the cross-correlation of every microphone pair is accumulated for each exploration region using

the delays pre-computed in Eq.1. In this way, we obtain a sound map at every time instant, as depicted in Figure 5. 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:

$$\hat{x} = \underset{x}{argmax} \sum_{i,j \in \Psi} R_{i,j}(\tau_{x,i,j}) \qquad (3)$$

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 estimation.

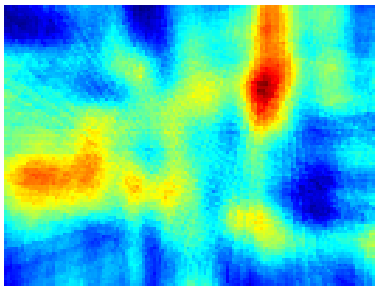


Figure 5 – Example of sound map obtained with the SRP PHAT process.

In the proposed scenario the acoustic localization system has to detect up to 2 acoustic sources produced simultaneously. We employ the method that dynamically estimates the number of sources based on a birth/death system. The ASL system uses a spatial segmentation algorithm to group locations that are close to each other in space and time. When a minimum number of locations N_b are found in a space region over a defined time window T_b, the system decides whether it is a new acoustic source. Similarly, if the previously detected acoustic source does not have any measurements that fall within its acceptance region for a given amount of time T_d, then it is dropped. The ratio between T_b and N_b used in the detection module is a design parameter. It must be high enough to filter out noises and outliers, but also not too high in order to be able to detect sporadic acoustic events. In our experiments N_b is set to 4, T_b is *460* ms and T_d is also *460* ms.

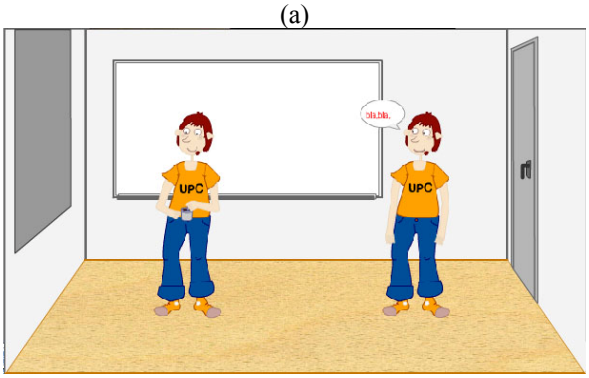
3.3 Visualization

We developed a graphical interface (GUI) that fully describes the acoustic activity in a smart room, and allows the observers to evaluate the system performance in a very convenient way. The GUI application is based on the QT Trolltech toolkit [15], an open-source (GPL) library widely used for the development of GUI programs. There are two screens in the GUI output, as shown in Figure 6. One corresponds to the real video captured from one of the cameras installed in the UPC's smart-room, and the other is a graphical representation of the output of the AED and ASL technologies.

When there is acoustic activity in the room, the GUI displays the animated puppets in the positions provided by the ASL system. The number of puppets depends on whether the acoustic event is produced in isolated manner (one puppet) or it is overlapped with speech (two puppets, one of which is always producing speech, as depicted in Figure 6).

Using this application, a video recording has been captured that contains the output of the GUI during a session lasting about 2 min, where three people in the room speak, interact with each other or produce one of the 12 (isolated as well as overlapped with speech) meeting-room AEs reported in Table 1. The video recording has not been edited, so it shows what can be seen in the room in real time. The two functionalities are simply juxtaposed in the GUI, so e.g. it may happen that the AED output is correct but the output of acoustic source localization is not, so showing the right event in a wrong place.

(a)



(b)

Figure 6 – The two screens of the GUI: (a) real-time video, and (b) graphical representation of the AED and ASL functionalities (''cup clink" overlapped with speech is being produced).

4. EXPERIMENTAL RESULTS

In order to prove the adequateness of the proposed approach, a series of experiments has been conducted to compare the implemented AED system working online with the baseline offline system [5] and the results are presented in Table 2. In both online and offline tests the first column corresponds to the detection accuracy of the isolated acoustic events and the second one corresponds to AEs overlapped with speech.

In our experiments we used 8 sessions of isolated acoustic events from the database described in Section 2. Additionally, these sessions were artificially overlapped with speech with different SNRs: -10 dB, 0 dB and +10 dB. For both offline and online tests, seven sessions (from 2 to 8) were used for training, and the remaining session 1 for testing.

The main difference between the online and the offline tests is in the way of processing the input waveform. During the offline tests the entire session is available for Viterbi segmentation. In this case the only parameter for tuning is the word insertion penalty parameter (*p*-value) that is a kind of trade-off between misses and false alarms. In our experiments p = -200. In online tests, the recognition is performed on a frame-by-frame basis using the additional modules described in Sub-section 3.2: silence detector, finite state machine, etc. In that case, more parameters have to be tuned: the silence threshold, the number of silence frames, the confidence thresholds for each AE, etc. Note that in the online tests the output hypothesis labels are those that are displayed by the visualization block.

Table 2 – Comparison of the recognition results (in percentage) between offline and online AED systems.

AEs	Offline system		Online system	
	Isolated	Overlap	Isolated	Overlap
ap	100	100	92	84
cl	100	100	85	89
cm	97	97	64	65
co	67	95	87	75
ds	83	100	84	84
kj	100	100	97	93
kn	100	95	52	72
kt	67	100	85	86
pr	100	96	92	97
pw	64	86	74	73
st	80	82	75	70
Average	**91.3 %**		**80.6%**	

As can be seen from Table 2, almost all AEs are well detected in offline simulations. Relatively low detection rate corresponds to low-energy AEs, such as "keyboard typing", "paper work" and "steps"; additionally, the AE "cough" is often confused with speech. In online simulations the best detection rate is achieved for AEs "applause", "cup clink", "key jingle" and "phone ring"."Door knock" and "chair moving" showed relatively low detection rates; actually, these AEs have the shortest duration in the employed testing database. In average, 80.6% of accuracy is achieved in online simulations.

5. CONCLUSIONS AND FUTURE WORK

In this work we developed a 2-source acoustic event detection and localization system running in real-time in the UPC's smart-room. The detection of AEs is performed using a HMM approach, which allows analyzing the input waveform on a frame-by-frame basis that offers low latency. The AED and ASL systems are visually monitored by a GUI application which shows the output of AED and ASL technologies jointly in real-time.

In order to visualize the 2 acoustic sources of the overlapped AE in its correct position in the room, we adopted a scenario where speech can only appear at the right part and the remaining AE at the left. To remove this constraint, future work will be devoted to developing source separation techniques in the room using multi-microphone processing.

6. ACKNOWLEDGEMENTS

This work has been funded by the Spanish project SARAI (TEC2010-21040-C02-01). The authors are grateful to Eros Blanco, Andrey Temko and Joan-Isaac Biel, co-authors of the previous versions of the system, for their contribution in the development of SmartFlow-based software deployed in the Smart-room. The first author is partially supported by a grant from the Catalan autonomous government.

REFERENCES

[1] CLEAR, 2006. Classification of Events, Activities and Relationships. Evaluation and Workshop. <http://isl.ira.uka.de/clear06>.

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

[3] A. Temko, C. Nadeu, J.-I. Biel, 2008, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07", *Multimodal Technologies for Perception of Humans*, LNCS, v. 4625, Springer. pp. 354–363, 2008.

[4] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson and T. S. Huang, "Real-world acoustic event detection", *Pattern Recognition Letters*, vol. 31, issue 12, pp. 1543-1551, 2010.

[5] T. Butko, C. Canton-Ferrer, C. Segura, X. Giro, C. Nadeu, J. Hernando, J.R. Casas, "Improving detection of acoustic events using audiovisual data and feature level fusion", *Proc. Interspeech*, 2009.

[6] A. Temko, C. Nadeu, "Acoustic event detection in meeting-room environments", *Pattern Recognition Letters*, vol. 30/14, pp 1281-1288, Elsevier, 2009.

[7] A. Waibel and R. Stiefelhagen, *Computers in the human interaction loop*, Springer, New York, USA, 2009.

[8] E. Blanco, I*dentification of two simultaneous sources in a real meeting-room environment*, Master Thesis, Politecnico di Milano and UPC, 2009.

[9] C. Nadeu, D. Macho, J. Hernando, "Frequency & time filtering of filter-bank energies for robust HMM speech recognition", *Speech Communication*, vol. 34, pp. 93-114, 2001.

[10] J. Luque and J. Hernando, "Robust speaker identification for meetings: UPC CLEAR-07 meeting room evaluation system", *Multimodal Technologies for Perception of Humans*, LNCS, vol. 4625/2008, pp. 266-275, 2008.

[11] L. Rabiner, B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[12] S. Young, et al., *The HTK Book (for HTK Version 3.2)*", Cambridge University, 2002.

[13] J. Dibiase, H. Silverman, M. Brandstein, *Microphone Arrays. Robust Localization in Reverberant Rooms*, Springer, 2001.

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

[15] QT Trolltech toolkit , http://trolltech.com/products/qt.