CLEAR Evaluation of Acoustic Event Detection and Classification Systems

Andrey Temko[1], Robert Malkin[2], Christian Zieger[3],
Dusan Macho[1], Climent Nadeu[1], and Maurizio Omologo[3]

[1] TALP Research Center, UPC, Campus Nord, Ed. D5,
Jordi Girona 1-3, 08034 Barcelona, Spain
{temko, dusan, climent}@talp.upc.es
[2] interACT, Carnegie Mellon University,
407 S. Craig St, Pittsburgh PA 15213 USA
rgmalkin@cs.cmu.edu
[3] ITC-irst, via Sommarive 18, 38050, Povo (TN), Italy
{zieger, omologo}@itc.it

Abstract. In this paper, we present the results of the Acoustic Event Detection (AED) and Classification (AEC) evaluations carried out in February 2006 by the three participant partners from the CHIL project. The primary evaluation task was AED of the testing portions of the isolated sound databases and seminar recordings produced in CHIL. Additionally, a secondary AEC evaluation task was designed using only the isolated sound databases. The set of meeting-room acoustic event classes and the metrics were agreed by the three partners and ELDA was in charge of the scoring task. In this paper, the various systems for the tasks of AED and AEC and their results are presented.

1 Introduction

Although speech is certainly the most informative acoustic event, other kind of sounds may also carry useful information in a meeting room environment. In fact, in that environment the human activity is reflected in a rich variety of acoustic events, either produced by the human body or by objects handled by humans. Consequently, detection or classification of acoustic events may help to detect and describe the human and social activity that takes place in the room. For example: clapping or laughter inside a speech discourse, a strong yawn in the middle of a lecture, a chair moving or door noise when the meeting has just started, etc Additionally, the robustness of automatic speech recognition systems may be increased by a previous detection of the non-speech sounds lying in the captured signals.

Acoustic Event Detection/Classification (AED/C) is a recent sub-area of computational auditory scene analysis [1] that deals with processing acoustic signals and converting them into symbolic descriptions corresponding to a listener's perception of the different sound events that are present in the signals and their sources. While acoustic event classification deals with events that have already been isolated from its temporal context, acoustic event detection refers to both identification and localization in time of events in continuous audio streams.

R. Stiefelhagen and J. Garofolo (Eds.): CLEAR 2006, LNCS 4122, pp. 311 – 322, 2007.

In this paper, we present the results of the AED/C CLEAR evaluations carried out in February 2006 by the three participant partners from the CHIL project [2] which sign this paper (UPC, CMU and ITC). The primary evaluation task was AED of the testing portions of the two isolated sound databases (from ITC and UPC) and 4 UPC's seminar recordings produced in CHIL. Additionally, a secondary AEC evaluation task was designed using only the isolated sound databases, and it is also included in this report. All the partners agreed the set of acoustic classes a priori before recording the databases. A common metrics was also developed at the UPC and agreed with the other partners. ELDA was in charge of the scoring task. In this paper, the three participant sites present their own preliminary systems for the tasks of AED and AEC. Two of them are based on the classical Hidden Markov Model (HMM) [3] approach used in continuous speech recognition, and the other uses Support Vector Machine (SVM) [4] as the basic classifier. Since the evaluation procedure was not strictly defined, there are some differences between the degrees of fitting of the systems to the testing data: two partners developed specific systems for each room, but not the third; one partner uses a system trained differently for seminars and isolated event databases, etc. If those differences are neglected, it is observed that the system closest to the usual speech recognition approach offers better average AED results.

The paper is organized as follows: Section 2 gives the experimental setup. Specifically, the databases used in the evaluations are described in Subsection 2.1, while the evaluation scenario and metrics are given in Subsection 2.2 and 2.3, respectively. Section 3 reviews the systems used by each of the AED/C evaluation participants. The results obtained by the detection and classification systems in the CLEAR evaluations are shown and discussed in Section 4. Conclusions are presented in Section 5.

2 Evaluation Setup

2.1 Databases

The conducted experiments were carried out on 2 different kinds of databases, namely: 2 databases of isolated acoustic events recorded at the UPC and IRST, and 5 interactive seminars recorded at the UPC.

The two former databases contain a set of isolated acoustic events that occur in a meeting room environment and were recorded specially for the CHIL AED/C task. The recorded sounds do not have temporal overlapping and no interfering noises were present in the room.

The UPC database of isolated acoustic events [5] was recorded using 84 microphones, namely, Mark III (array of 64 microphones), three T-shape clusters (4 mics per cluster), 4 tabletop directional and 4 omni-directional microphones. The database consists of 13 semantic classes plus "unknown". Approximately 60 sounds per each of the sound classes were recorded as shown in Table 1. Ten people participated in recordings: 5 men and 5 women. There are 3 sessions per each participant. At each session, the participant took a different place in the room out of 7 fixed different positions.

The ITC database of isolated acoustic events [6] was recorded with 32 microphones. They were mounted in 7 T-shaped arrays (composed by 4 microphones each one) plus there were 4 table microphones. The database contains 16 semantic classes of events. Approximately 50 sounds per almost each of the sound classes were recorded as shown

in Table 1. 9 people participated at the recordings. For each experiment 4 positions in the room were located. People swapped their positions after every session. During each session every person reproduced a complete set of acoustic events.

Additionally, the AED techniques were applied to the database of the interactive seminars [7] recorded at the UPC. 5 interactive seminars have been collected. The difference with two previous databases of isolated acoustic events is that seminars consist of real environment events that may have temporal overlapping with speech and/or other acoustic events. Each seminar consists of a 10-20 minutes presentation to a group of 3-5 attendees in a meeting room. During and after the presentation there are questions from the attendees with answers from the presenter. There is also activity in terms of people entering/leaving the room, opening and closing the door, standing up and going to the screen, some discussion among the attendees, coffee breaks, etc. The databases was recorded using 88 different sensors that include 3 4-microphoneT-shaped arrays, 1 64-microphone Mark III array, 4 omni-directional table-top microphones, 4 directional table-top microphones, and 4 close-talk microphones. The number of events of one of the seminars is summarized in Table 1.

Table 1. Number of events for the UPC and ITC databases of isolated acoustic events, and the UPC interactive seminar

Event type	Number of events		
	UPC-isolated	ITC-isolated	UPC-seminar
Door knock	50	47	4
Door open	60	49	7
Door slam	61	51	7
Steps	73	50	43
Chair moving	76	47	26
Spoon/cup jingle	64	48	15
Paper work	84	48	21
Key jingle	65	48	2
Keyboard typing	66	48	14
Phone ring	116	89	6
Applause	60	12	2
Cough	65	48	5
Laugh	64	48	8
Unknown	126		12
Mimo pen buzz		48	
Falling object		48	
Phone vibration		13	
Speech			169

2.2 Evaluation Scenario

The AED/C evaluation is done on 12 semantic classes that are defined as:

- Knock (door, table) [kn]
- Door slam [ds]

- Steps [st]
- Chair moving [cm]
- Spoon (cup jingle) [cl]
- Paper wrapping [pw]
- Key jingle [kj]
- Keyboard typing [kt]
- Phone ringing/Music [pr]
- Applause [ap]
- Cough [co]
- Laugh [la]

Also there are two other possible events that are present but are not evaluated

- Speech [sp]
- Unknown [un]

Actually, the databases of isolated acoustic events contain more semantic classes than the above-proposed list as shown in Table 1. For that reason, the classes that are out of the scope of the current AED/C evaluation were marked as "unknown".

Two main series of experiments are performed: AED and AEC. AED was done in both isolated and real environment conditions. For the task of AEC and isolated AED the databases of isolated acoustic events were split into training and testing parts, namely, for the UPC database sessions 1 and 2 were used for training and session 3 for testing; for the ITC database sessions 1-3 were used for training and session 4 for testing. For the task of AED in real environment all databases of isolated acoustic events and one of five seminars were allowed to use for training and developing, while for testing a 5-minute extract from each of the remaining 4 seminars was proposed forming in total 4 five-minute segments. The selection of extracted parts was done by ELDA.

The primary evaluation task was defined as AED evaluated on both the isolated databases and the seminars.

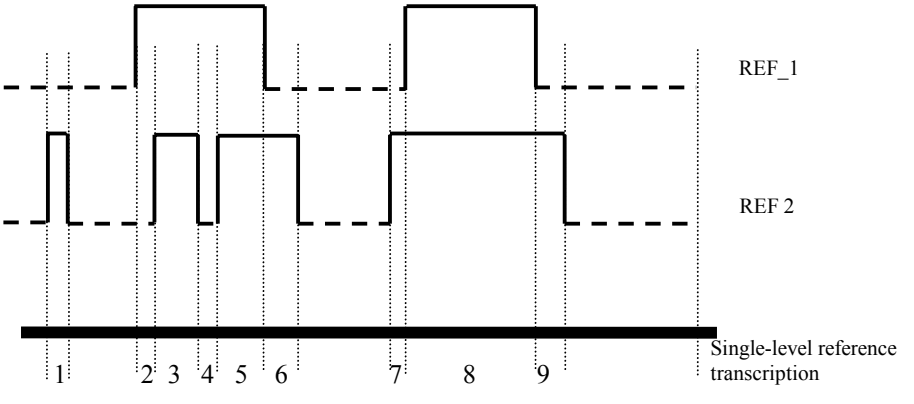

Fig. 1. From reference transcription with overlapping of level 2 to reference single-level transcription

Table 2. Obtained single-level reference transcription and a list of events to detect

Single-level reference *transcription*	*List of* ***events to detect****:*
1 – co1	1 – cough1
2 – la1	2 – laugh1
3 – la1_ds1	3 – ds1
4 – la1	4 – spoon1
5 – la1_cl1	5 – laugh2
6 – cl1	6 – keyboard1
7 – la2	

2.3 Metrics

As it was mentioned above, the acoustic events that happen in real environment may have temporal overlapping. The appropriate metric was developed to score the system outputs. It consists of two steps: projecting all levels of overlapping events into a single-level reference transcription and comparing a hypothesized transcription with the single level reference transcription.

For instance, let's suppose we have a reference that contain overlapping of level 2 and can be represented as shown in Figure 1 and

REF_1: _la_kt_
REF_2: _co_ds_cl_la_

where REF_1 and REF_2 model two overlapping acoustic event sequences. Then we can form the single-level reference transcription and a list of events to detect as shown in Table 2.

Following definitions are needed to compute the metric:

- An event is **correctly detected** when the hypothesised temporal centre is situated in the appropriate *single-level reference* interval and the hypothesised label is a constituent or a full name of this interval *single-level reference* label. After an event is claimed to be correctly detected, it is marked as detected in the list of *events to detect.*
- **Empty intervals** are the reference intervals that contain speech, silence or events belonging to the "unknown" class.
- **A substitution error** occurs when the temporal centre of the hypothesised event is situated in the appropriate *single-level reference* interval and the label of the hypothesised event is ***not*** constituent or the full name of the label of that *single-level reference* interval.
- **An insertion error** occurs when the temporal centre of the hypothesised event is ***not*** situated in any of the *single-level reference* intervals (i.e. are situated in *empty intervals*)
- A **deletion error** occurs when there is an event in the list of *events to detect* that is not marked as detected.

Finally, Acoustic Event Error Rate (AEER) is computed as

$$\mathbf{AEER = (D+I+S)/N * 100}$$

where N is the number of *events to detect*, D – *deletions*, I – *insertions*, and S – *substitutions*.

3 Acoustic Event Detection and Classification Systems

3.1 UPC AED/C Systems

A system based on SVM was used at the UPC for the task of AED/C. A DAG [8] multi-classification scheme was chosen to extend the SVM binary classifier to the multi-classification problem. 5-fold cross-validation [4] on the training data was applied to find the optimal SVM hyper parameters that were σ for the chosen Gaussian kernel, and C, a parameter that controls the amount of data allowed to be misclassified during the training procedure. In all the experiments the third channel of the Mark III microphone array was used.

Firstly, the sound is downsampled from the initial 44kHz sampling rate to 22 kHz, and framed (frame length=25ms, overlapping 50%, Hamming window). For each frame, the set of spectral parameters that showed the best results in [9] was extracted. It consists of the concatenation of two types of parameters: 1) 16 Frequency-Filtered (FF) log filter-bank energies [10] taken from ASR, and 2) a set of other perceptual parameters: zero-crossing rate, short time energy, 4 subband energies, spectral flux

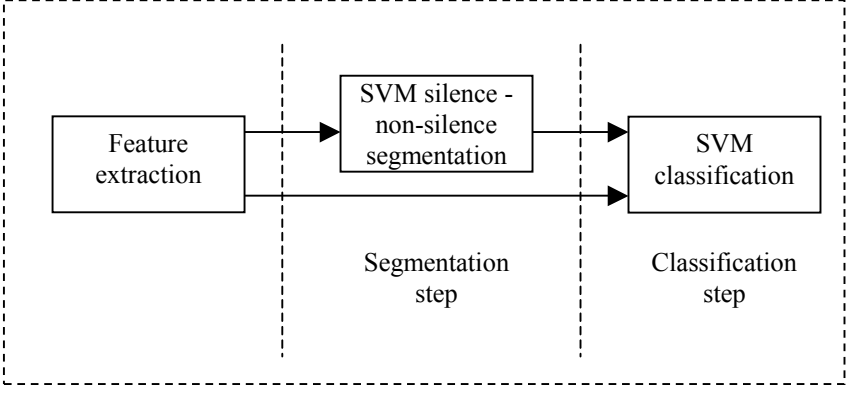

Fig. 2. UPC acoustic event detection system

calculated for each of the defined subbands, and pitch. The first and second time derivatives were also calculated for the FF parameters. In total, a vector of 59 components is build to represent each frame.

AEC system

The mean, standard deviation, entropy and autocorrelation coefficient of the parameter vectors were computed along the whole event signal thus forming one vector per

audio event with 4x59 elements. Then, that vector of statistical features was used to feed the SVM classifier, which was trained on the training set of the two databases of isolated acoustic events. The resulting system, herewith named "UPC-C", was used to test both UPC and ITC databases of isolated acoustic events, so neither feature nor system adaptation related to a specific database was applied

AED system

The scheme of the AED system herewith named "UPC-D" is shown in Figure 2. Using a sliding window of one second with a 100ms shift, a vector of 4x59 statistical features was extracted like in the AEC system described in the last sub-section for each position of the window (every 100ms).

The statistical feature vector is then fed to an SVM-based silence/non-silence classifier trained on silence and non-silence segments of the two isolated acoustic events databases. At the output, a binary sequence of decisions is obtained. A median-filter of size 17 is applied to eliminate too short silences or non-silences.

Then, the SVM-based event classifier is applied to each detected non-silence segment. The event classifier was trained on a parameters extracted from a sliding window with 100ms shift applied to each event in the way that the first and the last windows still include more than 50% of the event content. The event classifier is trained on both isolated acoustic events and seminar databases to classify a set of 12 defined acoustical classes, plus classes "speech" and "unknown". A sequence of decisions made on a 1-second window every 100ms is obtained within the non-silence segment. That sequence is smoothed by assigning to the current decision point the label that is most frequent in a string of five decision points around the current one. Also, a confidence measure is calculated for each point as the quotient between the number of times that the chosen label appears in the string and the number of labels in the string (5).

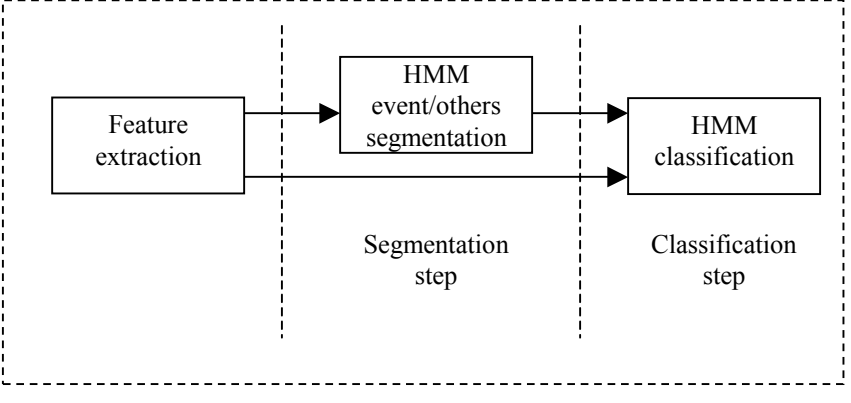

Fig. 3. CMU acoustic event detection system

The sequence of decisions from the non-silence segment is then processed again to get the detected events. In that step, only the events that have their length equal or larger than the average event length are kept, and the number of events kept in the non-silence segment is forced to be lower than a number which is proportional to the

length of the segment. The average length of the events is estimated from the training and development databases. Finally, if the average of the above mentioned computed confidences in a detected event is less than a threshold, the hypothesized event is marked as "unknown"; otherwise, it maintains the assigned label.

3.2 CMU AED/C Systems

The CMU acoustic event classification and detection systems were based on continuous density HMMs. We first downsampled the input signal from a single microphone to 16kHz, 2-byte quality. From this signal, we extracted 15 Mel-Frequency Cepstral Coefficients (MFCCs) at a rate of 100 frames per second. We additionally normalized these MFCCs to zero mean and unity variance using means and variances specific to each site. We used custom HMM topologies for each sound class; these topologies were induced using the k-variable k-means algorithm due to Reyes-Gomez and Ellis [11]. The k-variable k-means algorithm is a greedy approach to topology induction based on the leader-follower clustering paradigm; it uses a threshold to control the tendency to add new states to a class HMM.

We trained five complete sets of class HMMs using all available data from the isolated databases. After training these five complete HMM sets, we further trained site-specific feature space adaptation matrices that are reflected in systems "CMU-C1" and "CMU-C2". We used the maximum likelihood approach suggested by Leggetter and Woodland [12] and Gales [13]. Finally, as suggested by Reyes-Gomez and Ellis, we explored the combination of scores of HMMs trained with different thresholds on a per-site basis. We found that by combining three models for the ITC data and two for the UPC data, we were able to achieve a combined misclassification rate of less than 6% for acoustic event classification task.

For the acoustic event detection task, we wished to explore the possibility of pre-segmenting the data with a simple HMM before applying our more complex classification HMMs which used more than one Viterbi path to assign a final score. The scheme of the system is presented in Figure 3. Hence, we trained segmentation HMMs which included three classes: speech, CHIL event, and other. To train these HMMs, we used the same approach as for the classification systems above, except that we added the UPC seminar data for training. The detection systems herewith will be named "CMU-D1" and "CMU-D2". We chose the optimal HMMs for segmentation on a per-site basis. Further, since we also needed to control the rate at which these HMMs created segments in the data, we optimized separate insertion penalties for the ITC isolated database, the UPC isolated database, and the UPC seminar database. This approach yielded poor results on the isolated condition, and very poor results for the seminar condition.

3.3 ITC AED/C Systems

The AED/C system that was studied at the ITC-irst is based on continuous density HMM. The scheme of the system is presented in Figure 4.

A signal acquired by a single microphone belonging to a T-shaped array was used in experiments. The front-end processing is based on 12 Mel-Frequency Cepstral Coefficients (MFCCs) [3] and log-energy of the signal. The analysis step is 10 ms

with a Hamming window of 20 ms. The resulting parameters together with their first and second order time derivatives are arranged into a single observation vector of 39 components.

Each event is described by a 3 state HMM model. All of the HMMs have a left-to-right topology and use output probability densities represented by means of 32 Gaussian components with diagonal covariance matrices. HMM training was accomplished through the standard Baum-Welch training procedure.

For AEC task two different sets of models were created to fit the ITC and UPC rooms; the corresponding systems will herewith be named "ITC-C1" and "ITC-C2". The first one is trained on the ITC isolated acoustic events database and the other is trained on the UPC isolated acoustic events database. The selected training data refers to the recordings of a single microphone belonging to a T-shaped array.

For the AED task the same models adopted in the AEC task were used, but also the models of speech and silence were added; the corresponding systems will herewith be named "ITC-D1" and "ITC-D2". To train the model of speech, recordings of meetings in the ITC room were used, while for the model of silence the database of isolated acoustic events was used exploiting the silence moments between each event.

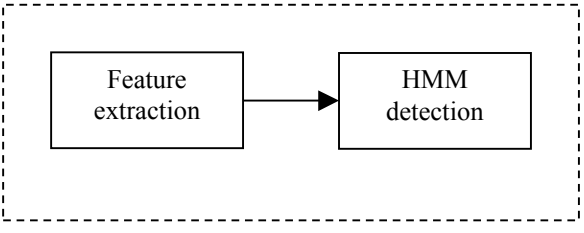

Fig. 4. ITC acoustic event detection system

To overcome the detection of events that are overlapped with speech that occur in the interactive seminars, the strategy based on the contamination of the events with speech in the training procedure was exploited; that is reflected in the system "ITC-D3". An artificial database was created by adding speech to the isolated events imposing different SNR values, from 0 to 15 dB. At this moment the system is not trained to detect events that overlap with other events except speech.

4 Results and Discussion

Table 3 shows classification error rates obtained using different classification systems described previously. Since the evaluation procedure was not strictly defined, there are some differences in the degree of fitting of the systems to the two testing databases (ITC and UPC isolated DB): both CMU and ITC systems use two sets of models, one for each testing database, while UPC system uses one set of models for the both testing databases. We can observe that the system based on SVM obtained the same or better results than the systems based on the HMM technology, despite the fact that database-specific systems were used in the case of HMM.

Table 3. Error rates (in %) for AE classification task of the systems explained in Section 3

Systems / Databases	UPC-C	CMU–C1	CMU–C2	ITC–C1	ITC–C2
ITC isolated DB	4.1	7.5	----	12.3	----
UPC isolated DB	5.8	----	5.8	----	6.2

In the detection task, as explained in the previous sections, participants took two different approaches: a) First performing segmentation and then classification (UPC and CMU systems) b) Merging the segmentation and classification in one step as performed by the Viterbi search in the state-of-the-art ASR systems (ITC systems) Table 4 shows detection error rates for the two isolated event databases and the interactive seminar database. The lowest detection error rates are obtained by the ITC systems, which are based on the approach b). Notice that both CMU and UPC systems achieved better results than the ITC systems in the classification task (Table 3), however they both rely on a previous segmentation step (the approach a)). If we add up the results obtained for the detection task for both isolated and seminar conditions neglecting the test-specificities of the CMU and ITC systems, we obtain the following error rates: UPC: 69.6%, CMU: 80.5%, ITC: 46.8%. Although there might be a number of reasons to explain the differences across the systems, we conjecture that the initial segmentation step included in both UPC and CMU systems, but not in the ITC systems, is the main cause of the lower overall detection performance of these systems. Further investigation is needed in the direction of the approach a) to see whether it can outperform the well-established scheme b).

Besides, it can be seen from the Table 4, that the error rates increase significantly for the UPC seminar database. One of possible reasons of such a bad performance is that it is difficult to detect low-energy acoustic classes that overlap with speech, such as e.g. “chair moving”, “steps”, “keyboard typing”, and “paper work”. Actually, these classes cover the majority of the events in the UPC seminars and probably they are the cause of the bad results we obtained in the seminar task. A usage of multiple microphones might be helpful in this case.

Table 4. Error rates (in %) for AE detection task of the systems explained in Section 3

Systems / Databases	UPC–D	CMU–D1	CMU-D2	ITC-D1	ITC-D2	ITC-D3
ITC isolated DB	64.6	45.2	----	23.6	----	----
UPC isolated DB	58.9	----	52.5	----	33.7	----
UPC seminars DB	97.1	----	177.3	----	----	99.3

5 Conclusions

The presented work focused on the CLEAR evaluation tasks concerning the detection and classification of acoustic events that may happen in a lecture/meeting room environment. In this context, we evaluated two different tasks, acoustic event classification (AEC) and acoustic event detection (AED), AED being the primary objective of

the evaluation. Two kinds of databases were used, two databases of isolated acoustic events and a database of interactive seminars containing a significant number of acoustic events of interest.

Preliminary detection and classification systems from three different participants were presented, which allowed an evaluation of different approaches for both classification and detection. The UPC system is based on the Support Vector Machine (SVM) discriminative approach and uses Frequency Filtering features and four kinds of perceptual features. Both the CMU and ITC systems are based on the Hidden Markov Model (HMM) generative approach and they use MFCC features.

In the classification task, the UPC SVM-based system showed better performance than the two systems based on HMM. In the detection task, we could see two different approaches: a) first performing segmentation and then classification (UPC and CMU systems), and b) merging the segmentation and classification in one step as performed by the Viterbi search in the state-of-the-art Automatic Speech Recognition (ASR) systems (ITC systems). In the presented results, the approach b) showed better performance than the approach a). Notice however that the b) approach (and actually the ITC systems) is a well-established ASR approach developed for many years and thus can be considered as a challenging reference for the other presented approaches/systems in the acoustic event detection task.

Acknowledgements

This work has been partially sponsored by the EC-funded project CHIL (IST-2002-506909). Authors wish to thank Djamel Mostefa and Nicolas Moreau from ELDA for their role in the transcription of the seminar data and in the scoring task.

UPC authors have been partially sponsored by the Spanish Government-funded project ACESCA (TIN2005-08852).

References

1. D. Wang, G. Brown, Computational Auditory Scene Analysis: Principles, Algorithms and Applications, Wiley-IEEE Press, 2006
2. CHIL - COMPUTERS IN THE HUMAN INTERACTION LOOP, http://chil.server.de/
3. L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993
4. B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002
5. A. Temko, D. Macho, C. Nadeu, C. Segura, "UPC-TALP Database of Isolated Acoustic Events", *Internal UPC report*, 2005
6. C. Zieger, M. Omologo, "Acoustic Event Detection - ITC-irst AED database", *Internal ITC report*, 2005
7. J. Casas, R. Stiefelhagen, et al, "Multi-camera/multi-microphone system design for continuous room monitoring," *CHIL-WP4-D4.1-V2.1-2004-07-08-CO, CHIL Consortium Deliverable D4.1*, July 2004
8. J. Platt et al., "Large Margin DAGs for Multiclass Classification", *Proc. Advances in Neural Information Processing Systems* 12, pp. 547-553, 2000
9. A. Temko, C. Nadeu, "Classification of meeting-room acoustic events with Support Vector Machines and Confusion-based Clustering", *Proc. ICASSP'05*, pp. 505-508, 2005

10. C. Nadeu et al., “On the decorrelation of filter-bank energies in speech recognition”, *Proc. Eurospeech’95,* pp. 1381-1384, 1995
11. M. Reyes-Gomez and D. Ellis, “Selection, Parameter Estimation, and Discriminative Training of Hidden Markov Models for General Audio Modeling”, *Proc. ICME’03*, 2003
12. C. Leggetter and P. Woodland, “Speaker Adaptation of Continuous Density HMMs using Multivariate Regression*”, Proc. ICSLP’94*, 1994
13. M. Gales, “Maximum Likelihood Linear Transformations for HMM-based Speech Recognition”, Computer Speech and Language, 1998