7

Acoustic Event Detection and Classification

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

[1] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
[2] interACT, Carnegie Mellon University, Pittsburgh, PA, USA
[3] FBK-irst, Foundation Bruno Kessler, Povo, Italy

The human activity that takes place in meeting rooms or classrooms is reflected in a rich variety of acoustic events (AE), produced either by the human body or by objects handled by humans, so the determination of both the identity of sounds and their position in time may help to detect and describe that human activity. Indeed, speech is usually the most informative sound, but other kinds of AEs may also carry useful information, for example, clapping or laughing inside a speech, a strong yawn in the middle of a lecture, a chair moving or a door slam when the meeting has just started. Additionally, detection and classification of sounds other than speech may be useful to enhance the robustness of speech technologies like automatic speech recognition.

Acoustic event detection and classification (AED/C) is a recent discipline that may be included in the broad area of computational auditory scene analysis [1]. It consists of processing acoustic signals and converting them into symbolic descriptions corresponding to a listener's perception of the different sound events present in the signals and their sources.

This chapter presents the evolution of the pioneering works that start from the very first CHIL dry-run evaluation in 2004 and end with the second international CLEAR evaluation campaign in 2007. During those years of the CHIL project, the task was significantly modified, going from the classification of presegmented AEs to the detection and classification in real seminar conditions where AEs may have temporal overlaps. The changes undergone by the task definition along with a short description of the developed systems and their results are presented and discussed in this chapter.

It is worth mentioning several factors that make the tasks of AED/C in meeting-room environments particularly challenging. First, detection and classification of sounds have been usually carried out so far with a limited number of classes, e.g., speech and music, but there were no previously published works about meeting-room AEs when the CHIL project was launched. Consequently, the sound classes and what is considered an AE instance had to be defined. Second, the nature of AEs is different from speech, and features describing speech's spectral structure are not necessarily suitable for AED/C, so the choice of appropriate features becomes an

important issue. Third, the chosen classifiers have to face the data scarcity problem caused by the lack of large specific databases. Fourth, the low signal-to-noise ratio of the recorded AEs signals is responsible for a high error rate. It is due to both the use of far-field microphones (required by the unobtrusiveness condition) and the fact that the criterion of natural interaction produced a lot of temporal overlaps of AEs with co-occurring speech and other AEs. Fifth, the evaluation of those new AED/C tasks required the establishment of an (agreed on) protocol to develop databases and appropriate metrics. In particular, recordings had to find a balance between people's natural behavior and a large enough number of recorded AEs. And six, the evaluation data come from up to five different smart rooms with different acoustic conditions and microphone locations.

The chapter is organized as follows. We start with AEC as this task was first pursued in the CHIL project because it is easier than AED. Section 7.1 presents the work carried out on AEC within three evaluations organized by CHIL, namely, the dry-run evaluation in 2004, the CHIL international evaluation in 2005, and the CLEAR international evaluation campaign in 2006 (see also Chapter 15). The definition of a set of meeting-room AEs, the task, and metrics are given in this section. The developed AEC systems are briefly described and their results are presented and discussed. While AEC is performed on previously segmented AEs, AED is a more difficult task because the identity of sounds and their position in time have to be obtained. In Section 7.2, we overview the two international evaluation campaigns where the AED task was presented, namely, CLEAR 2006 [5] and CLEAR 2007 [4]. The AED task definition, metrics, and evaluation setup are described. A summary of the developed systems and their results are proposed. Section 7.3 overviews AED demonstrations developed by CHIL project partners. Conclusions and remaining challenges are presented in Section 7.4.

7.1 Acoustic Event Classification

On the way toward the detection of AEs, it was first decided to perform classification of AEs. Let us note that classification deals with events that have been already extracted or, alternatively, the temporal positions of acoustic events in an audio stream are assumed to be known. The AEC task has been evaluated in the dry-run evaluation in 2004, the first CHIL evaluation in 2005, and the CLEAR 2006 evaluation campaign. Specifically, Section 7.1.1 presents the classes and DBs used in the evaluations. Evaluation protocol, participants, and metrics are given in Section 7.1.2. The various approaches, their results, and a discussion are presented in Section 7.1.3.

7.1.1 Acoustic Event Classes and Databases

Table 7.1 shows the acoustic event classes considered in the three evaluations. In the very first dry-run evaluations, seminar data recorded at University of Karlsruhe (UKA-ISL), which were also used for other CHIL evaluations, were transcribed according to 36 sound classes. The instances were transcribed with tight temporal

bounds, allowing isolated-sound tests to be performed. In that corpus, 25 classes were found to be suitable for use in the evaluation based on the number of instances for both training and testing. That was an ad hoc set, in the sense that we did not know a priori what kinds of sounds would be present in the database, since the recordings were designed without planning the occurrence of acoustic events, and we relied on the transcriber's judgment to select the whole sound set. This resulted in a large number of class overlaps (e.g., "bang" and "bump"), but it allowed us to determine which classes were actually present and their relative number of instances. In this way, 2805 segments were taken in total and divided into training (1425 segments) and evaluation sets (1380) at random.

Due to the low performance shown in the dry-run evaluations and in order to limit the number of classes to those that were both easy to identify for transcribers and semantically relevant to the CHIL task, the number of classes for the CHIL evaluations in 2005 was decreased from 25 to 15. A few semantically similar classes were combined; e.g., the three classes "breath", "inspiration", and "expiration" were grouped into one class "breath", as can be seen in Table 7.1. Additionally, several new classes were added, like the music/phone ring class or the generic mouth noise class. However, as we wished to retain a label for every sound in the corpus, whether it was easily identifiable or not, we split the label set into two types: semantic labels and acoustic labels. Semantic labels corresponded to specific named sounds with CHIL relevance, e.g., door slamming, speech, etc. Acoustic labels corresponded to unnamed sounds that were either unidentifiable or had no CHIL relevance. Based on the availability of samples per class, 15 semantic classes were finally chosen as shown in Table 7.1. Table 7.2 shows the acoustic labels used in the CHIL evaluations in 2005. The acoustic labels used names describing both tonal and rhythmic features of a sound. For acoustic evaluation, all semantically labeled classes were mapped to the corresponding acoustic classes, as shown in Table 7.2. From seminars recorded at UKA-ISL, 7092 segments were extracted; 3039 were used for training (1946 "speech", 333 "breath", 333 "silence", 232 "unknown", etc.), 1104 for development (749 "speech", 183 "unknown", 139 "silence", 9 "cough", etc.), and 2949 for evaluation (2084 "speech", 348 "silence", 274 "unknown", 123 "breath", etc.).

Classes / **Year**	Breath	Expiration	Inspiration	Conversation	Speech	Cough	Throat	Applause	Bang	Bump	Click	Chair moving	Crump	Door slam	Door knock	Keyboard	Key jingle	Laughter	Metal	Microphone	Mouth noise	Music/phone ring	Movements	Paper work	Pen writing	Pop	Shh...	Silence	Smack	Snap	Sniff	Spoon-cup jingle	Squeak	Steps	Tap	Unknown	Whirring	**Total**
2004	√	√	√	√	√	√	√		√	√	√		√	√		√			√	√			√		√	√	√	√	√	√			√	√			**25**
2005		√		√		√					√	√		√		√					√	√		√			√			√			√		√	√	**15**
2006						√	√					√		√	√	√	√	√				√		√							√		√				**15**

Table 7.1. Evaluated acoustic event classes for evaluations that took place in years 2004, 2005, and 2006.

Semantic Label	Acoustic Label
laughter, speech	continuous tone
breath, sniff, paper, whirring, cough	continuous sound without tone
click, chair, door slam, mouth noise	single noncontinuous sound
steps	regular repetitive noncontinuous sound
keyboard typing	irregular repetitive noncontinuous sound
–	generic noise, all other kinds of sounds

Table 7.2. Semantic-to-acoustic mapping.

Due to the low number of instances for most semantic classes, it was not possible to obtain statistically significant results [2]. Besides, they were biased by prevailing classes like "speech" and "silence". For the CLEAR international evaluation campaign in 2006, we redefined the set of AE classes, including new ones that were judged as relevant for the meeting-room scenario, and reducing the number to 12 in order to make them more easily identifiable by transcribers. Besides, to get rid of the bias to "speech" and "silence" classes, the latter were discarded from evaluation, as Table 7.1 shows. To increase the number of instances for the chosen 12 classes, two partners, UPC and FBK-irst, recorded databases of isolated acoustic events [7]. From them each of 12 classes had about 100 instances, from which around two thirds were taken for training and the rest for testing.

7.1.2 Evaluation Protocol, Participants, and Metrics

In all evaluations, audio was collected with a combination of microphones. In the dry-run evaluations in 2004 and the CHIL evaluation in 2005, a channel of a wall-mounted 64-microphone array (Mark III array) was used, while in the CLEAR 2006 evaluation, any of around 90 far-field microphones (from the microphone array, several T-shaped clusters, and also tabletop microphones) was allowed to be used. Participants were permitted to use for system training only the data distributed for this evaluation and identified as training data. No other data were allowed for training purposes. Only an isolated sound test was performed. The sounds that were overlapped with other sounds were discarded. The Technical University of Catalonia (UPC) and Carnegie Mellon University (CMU) research groups participated in all three evaluations on the AEC. The Bruno Kessler Foundation (FBK-irst) group participated in the CLEAR 2006 evaluation. In the dry-run evaluations, accuracy and recall were used as metrics. The former is defined as the number of correctly classified acoustic events divided by the total number of acoustic events. Recall is defined as a mean of the individual class recalls calculated as the number of correctly classified events of a particular class divided by the number of events of that class. In the CHIL evaluations, the used error rate was defined as 100 minus accuracy (%). In the CLEAR 2006, a metric that counts the number of insertion, substitution, and deletion errors was introduced for the AED task; however, for the AEC, where only

substitution errors are possible, the same error rate defined for the CHIL evaluations in 2005 was used.

7.1.3 Systems, Results, and Discussion

The systems developed at CMU were based on HMM/GMM classifiers. Several HMM topologies were tried and the submitted system was obtained with a topology given by BIC for evaluations in 2004/2005 and by *k*-variable *k*-means algorithm for the CLEAR 2006 [2]. PCA was applied to a set of features composed of MFCC in 2004 and MFCC plus a few perceptual features in the 2005/2006 evaluations. For CLEAR 2006 evaluations, after training a complete set of HMMs, site-specific feature space adaptation matrices were further trained, resulting in two systems [7].

At UPC, after comparing the preliminary results obtained with GMM and SVM classifiers, the latter were finally chosen. A set of features composed of several perceptual features and frequency-filtered bank energies was exploited. Several statistical parameters were calculated from the frame-level features over all frames in an acoustic event. The obtained set of segment-level features was fed to the SVM classifier [7].

At FBK-irst, three state HMM models with MFCC features were used. All of the HMMs had a left-to-right topology and used output probability densities represented by means of 32 Gaussian components with diagonal covariance matrices. Two different sets of models were created to fit, respectively, the FBK-irst and UPC rooms; each set was trained with data from only one of the two isolated AE databases [7].

The results of the AEC evaluation are shown in Table 7.3. The low system performance obtained in 2004 was attributable to the relatively large number of classes. Additionally, the large number of classes made it rather difficult to correctly transcribe the seminars, resulting in a number of outliers. The scores from the accuracy measure are much higher than those from recall, showing that the results are biased by prevailing classes like “speech” and “silence”.

Evaluations	**Database**		**Metrics(%)**	**Baseline**	**CMU**		**UPC**	**FBK-irst**	
2004	Seminars 2004		Recall	4	26.4		24.15	–	
			Accuracy	47.8	61.6		55.1	–	
2005	Seminars 2004		Accuracy	47.8	63.2		62.9	–	
	Seminars 2005	Semantic	Error rate	29.4	27.2		23.5	–	
		Acoustic		29.3	26.7		27.8	–	
2006	Databases of isolated AEs	UPC	Error rate	–	7.5	–	4.1	12.3	–
		FBK		–	–	5.8	5.8	–	6.2

Table 7.3. Results of the AEC evaluations

The AEC 2005 evaluation unfortunately still suffered from problems of corpora. Transcription mistakes were still quite frequent; hence, the labels were not as reliable

as they could have been. Further, the corpus was extremely unbalanced; a large majority of the segments were made of speech. Many interesting classes had only a few examples in the training or test sets, and therefore they were not well-modeled by our systems. For the 2004/2005 evaluations, Table 7.3 shows additionally the baseline scores obtained as if a system just chose the most frequent class: "speech" in the 2004 evaluations; "speech" and "continuous tone" for semantic and acoustic sets, respectively, in the 2005 evaluations. Besides, Table 7.3 shows that systems submitted in 2005 improved the results obtained with the systems submitted in 2004 for the evaluation corpus used in 2004. For the 2006 evaluation, the error rate was smaller since the database collection had been designed specifically for acoustic events, and the number of instances per class was high enough and homogeneous among classes. The SVM-based system from UPC obtained the best error rate despite the fact that it was not database-specific.

7.2 Acoustic Event Detection

The AED task was evaluated in two international evaluation campaigns: CLEAR 2006 [5] and CLEAR 2007 [4]. In CLEAR 2006, we first focused on the easier task of detection of isolated AEs and made a preliminary attempt to perform AED in spontaneous conditions. In CLEAR 2007, only the task of spontaneous AED was tackled. Specifically, Section 7.2.1 presents the classes and DBs used in the evaluations. The evaluation setup is given in Section 7.2.2. The approaches, results, and discussion are presented in Section 7.2.3.

7.2.1 Acoustic Event Classes and Databases

In the AED 2006 and 2007 evaluations, the set of acoustic events used in the AEC 2006 evaluations was kept, which appears in Table 7.1. It is worth noting that, apart from the 12 evaluated acoustic classes, there are also three other classes that have not been evaluated though the present: "speech", "unknown", and "silence".

For CLEAR 2006 and CLEAR 2007, CHIL partners turned to recordings of scripted interactive seminars. Each seminar usually consists of a presentation of 10 to 30 minutes to a group of three to five 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, discussion among the attendees, coffee breaks, etc. Each meeting can be conditionally decomposed into acoustic scenes: "beginning", "meeting", "coffee break", "question/answers", and "end". The recorded interactive seminars contained a satisfactory number of acoustic events, so it was possible to perform AED tests that are statistically meaningful. The AEs are often overlapped with speech and/or other AEs.

In CLEAR 2006, two main series of experiments were performed: AED in the isolated condition and AED in the real environment condition. For the task of isolated AED, the databases of isolated acoustic events were split into training and testing

parts in the same way as it has been done for the AEC task. For the task of AED in real environments, five 20-minute seminars recorded at UPC were chosen as the richest ones in terms of the number of instances per class, from which one whole seminar was used for training along with all databases of isolated acoustic events, and four 5-minute extracts from the remaining four seminars were used for testing.

In CLEAR 2007, the seminars recorded at UKA-ISL, Foundation Bruno Kessler (FBK-irst), Athens Institute of Technology (AIT), and UPC were found suitable for evaluations, forming a set of 20 interactive seminars. Four interactive seminars (one from each site) were assigned for system development. Along with the seminar recordings, the whole databases of isolated AEs recorded at UPC and FBK-irst were used for development. In total, development seminar data consisted of 7495 seconds, where 16% of total time is AEs, 13% is silence, and 81% is "speech" and "unknown" classes.

From the remaining seminars, 20 5-minute segments were extracted for testing. In total, the test data consisted of 6001 seconds, where 36% is AE time, 11% is "silence", and 78% is "speech" and "unknown". Noticeably, about 64% of the time, the AEs are overlapped with "speech", and 3% of the times they overlap with other AEs.

7.2.2 Evaluation Protocol, Participants, and Metrics

The primary evaluation task in CLEAR 2006 was defined as AED evaluated on both the isolated databases and the seminars.

In order to have systems comparable across sites, a set of evaluation conditions was defined for CLEAR 2007: The evaluated system must be applied to the whole CLEAR 2007 test database; only primary systems are submitted to compete; the evaluated systems must use only audio signals, though they can use any number of microphones.

There were three participants in CLEAR 2006: CMU, UPC, and FBK-irst. In CLEAR 2007, however, eight sites signed up to participate; six sites submitted the results, while two withdrew their applications. The participating partners were AIT, Institute for Infocomm Research (IIR), FBK-irst, Tampere University of Technology (TUT), University of Illinois at Urbana-Champaign, (UIUC), and UPC. As mentioned above, the acoustic events that happen in a real environment may have temporal overlaps. In order to be able to properly score the output of the systems, appropriate metrics were developed by UPC and agreed upon by the other CHIL partners involved in the evaluations (FBK-irst, CMU, and ELDA).

For CLEAR 2006, the metric protocol consisted of two steps: (1) projecting all levels of overlapping events into a single-level reference transcription, and (2) comparing a hypothesized transcription with the single-level reference transcription. It is defined as the acoustic event error rate:

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

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

However, this metric had some disadvantages. The main assumption was that one reference acoustic event can cause only one error. In some cases, the metric was ambiguous, e.g., when one part of an acoustic event is detected correctly, another part of the same event causes a substitution error, and the rest is deleted, so the final score of the metric is affected by the last decision made about the acoustic event. For this purpose, for CLEAR 2007, this metric was decomposed into two other metrics: an F-score measure of detection accuracy (AED-ACC), and an error rate measure that focuses more on the accuracy of the endpoints of each detected AE (AED-ER).

The aim of the AED-ACC metric is to score the detection of all instances of what is considered as a relevant AE. With this metric, it is not important to reach a good temporal coincidence of the reference and system output timestamps of the AEs, but rather to detect their instances. It is oriented to applications like real-time services for smart rooms, audio-based surveillance, etc. The AED-ACC is defined as the F-score (the harmonic mean between precision and recall):

$$\text{AED-ACC} = \frac{(1+\beta^2) * \text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}},$$

where

$$\text{Precision} = \frac{\text{number of correct system output AEs}}{\text{number of all system output AEs}},$$

$$\text{Recall} = \frac{\text{number of correctly detected reference AEs}}{\text{number of all reference AEs}},$$

and β is a weighting factor that balances precision and recall. In this evaluation, the factor β was set to 1. A *system output (reference) AE* is considered *correct or correctly produced (correctly detected)* either (1) if there exists at least one reference (system output) AE whose temporal center is situated between the timestamps of the system output (reference) AE and the labels of the system output AE and the reference AE are the same, or (2) if the temporal center of the system output (reference) AE lies between the timestamps of at least one reference (system output) AE and the labels of the system output AE and the reference AE are the same.

For some applications, it is necessary to have a good temporal resolution of the detected AEs. The aim of this metric is to score AED as a task of general audio segmentation. Possible applications can be content-based audio indexing/retrieval, meeting-stage detection, etc. In order to define AED-ER, the NIST metric for speaker diarization was adapted to the task of AED. The audio data are divided into adjacent segments whose borders coincide with the points whether either a reference AE or a system output AE starts or stops, so that, along a given segment, the number of reference AEs and the number of system output AEs do not change. The AED-ER score is computed as the fraction of time, including regions of overlaps, in which a system output AE is not attributed correctly to a reference AE, in the following way:

$$(\text{AED - ER}) = \frac{\sum_{allseg.} dur(seg) * (max(N_{REF}, N_{SYS} - N_{correct}(seg)))}{\sum_{allseg.} dur(seg) * N_{REF}(seg)}$$

where, for each segment seg, dur(seg) is a duration of seg, N_{REF} (seg) is the number of reference AEs in seg, N_{SYS} (seg) is the number of system output AEs in seg, and $N_{correct}$ (seg) is the number of reference AEs in seg that correspond to system output AEs in seg. Notice that an overlapping region may contribute several errors.

7.2.3 Systems, Results, and Discussion

In CLEAR 2006, the systems used for AED were essentially those used for AEC. Two main directions were taken: (1) first performing segmentation and then classification (in the UPC and CMU systems), and (2) merging the segmentation and classification in one step as usually performed by the Viterbi search in the current state-of-the-art ASR systems (in the FBK-irst system). Specifically, the CMU system includes an HMM-based events/others segmentation before feeding the HMM event classifier. At UPC, silence/nonsilence segmentation was done on a sliding window of 1s, and the nonsilence portions were then fed to the SVM classifier with a subsequent smoothing and postprocessing. At FBK-irst, however, a standard ASR system was used. All systems submitted to CLEAR 2006 used a single-microphone channel, although CMU and UPC used a Mark III channel, while FBK-irst used a channel taken from a T-shaped microphone cluster.

In CLEAR 2007, the UPC system used in CLEAR 2006 was modified by eliminating the segmentation step, performing signal normalization, and using voting-based fusion of the decisions taken from four microphones. The FBK-irst systems remained almost the same from CLEAR 2006, only changing the number of Gaussians from 32 to 128. Site adaptation was also used in the FBK-irst system in CLEAR 2007. The system submitted by AIT used a hierarchical approach for event classification based on an HMM classifier with different HMM topologies per class and also building a specific system for each site. Signals taken from five microphones were averaged in order to cope with the ambient noise. IIR used a system based on HMM/GMM classifiers. For the AED task, a multichannel system based on HMM was implemented at TUT. Observation probabilities were calculated separately for one channel from each T-shaped microphone array. After that, all the observation probabilities were combined and the optimal path through all models was decoded. An HMM-based system with lattice rescoring using features selected by AdaBoost was implemented at UIUC. Only data from one microphone in the development seminar data were used in the system.

As a summary, five participants (FBK-irst, IIR, AIT, TUT, and UIUC) exploited HMM-based systems, and UPC used an SVM-based system. Three systems were multimicrophone (AIT / TUT / UPC), and three were single-microphone (FBK-irst/IIR/UIUC).

Table 7.4 shows the results obtained for the AED task in the CLEAR 2006 and CLEAR 2007 evaluations. As can be seen from the table, the lowest detection error rates for CLEAR 2006 [7] were obtained by the FBK-irst systems, which did not use the segmentation step. Notice that both the CMU and UPC systems achieved better results than the FBK-irst systems in the classification task (Table 7.3). Although a

number of reasons might explain the differences across the systems, we conjectured that the initial segmentation step included in both the UPC and CMU systems, but not in the FBK-irst systems, was the main cause of those systems' lower overall detection performance. Further investigation would be needed in that direction. Besides, as can be seen in Table 7.4, in the 2006 evaluations, the error rates increased significantly for the UPC seminar database. One possible reason for such a poor performance was the difficulty in detecting low-energy acoustic classes that overlap with speech, such as "chair moving", "steps", "keyboard typing", and "paper work". Actually, these classes cover the majority of the events in the UPC seminars and probably were the cause of the poor results obtained in the seminar task. Using multiple microphones might be helpful in this case.

Evals	Database		Metrics (%)	CMU		UPC	FBK-irst				AIT				IIR	TUT	UIUC
2006	Databases of IAE seminars	UPC	AEER	45.2	–	58.9	23.6	–	–		–				–	–	–
		FBK		–	52.5	64.6	–	33.7	–		–				–	–	–
		UPC		–	177.3	97.1	–	–	99.3		–				–	–	–
	Overall			80.5		69.6	46.8				–				–	–	–
2007	Seminars	AIT	AED-ACC	–		18.6	16.8	–	–	–	2	–	–	–	19.4	15.7	27.8
		FBK		–		25.1	–	30	–	–	–	3.9	–	–	22	20	37.3
		UKA		–		23.3	–	–	11.8	–	–	–	8.5	–	16.7	4	35.4
		UPC		–		23.9	–	–	–	29	–	–	–	3.5	30.8	19.4	42
	Overall			–		23	23.4*				4.4*				22.9	14.7	33.6
	Seminars	AIT	AED-ER	–		128	103	–	–	–	184	–	–	–	238	105	99
		FBK		–		157	–	87	–	–	–	155	–	–	161	92	99
		UKA		–		147	–	–	157	–	–	–	212	–	142	299	99
		UPC		–		120	–	–	–	103	–	–	–	246	145	105	100
	Overall			–		137	109*				203*				170	139	99

Table 7.4. Results of the AED evaluations. * means a different system is used for each site.

In CLEAR 2007 evaluation results [4], even the best scores (those from UIUC's) are very low: only 36.3% of accuracy, and almost 100% of error rate, due to temporal overlaps. The system with the best scores shows a relatively low error in the "steps" class, which accounts for 40% of the total number of AEs. On average, more than 71% of all error time occurs in overlapped segments. If they were not scored, the error rate of most submitted primary systems would be around 30-40%. These results indicate that there is still much room for system improvement in meeting-room spontaneous AED. The choice of the system structure may be important. For instance, none of the AED systems presented to the evaluations was built as a set of isolated detectors. This approach can be addressed in the future [6]. An improvement in performance can be expected from the use of several modalities. Some sounds can be detected by video technologies; e.g., "steps" can be assumed when detecting from the video an object moving around the smart room, etc.

7.3 Demonstrations of Acoustic Event Detection

7.3.1 FBK-irst Acoustic Event Detection and Classification Demo

This section describes the acoustic event detection and classification demo developed by FBK-irst under the CHIL project. First, the implemented system detects the acoustic event through a detector, that is based on spectral variation functions. Then it derives its position using a localizer based on the global coherence field (GCF). Finally, it identifies the type, choosing among all possible CHIL events through an HMM-based classifier. The video of the demo has been recorded[1].

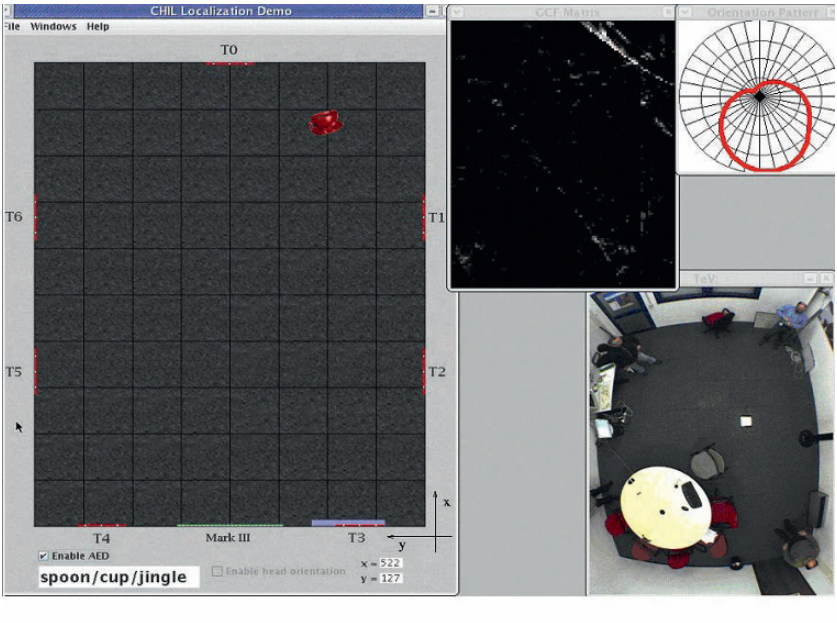


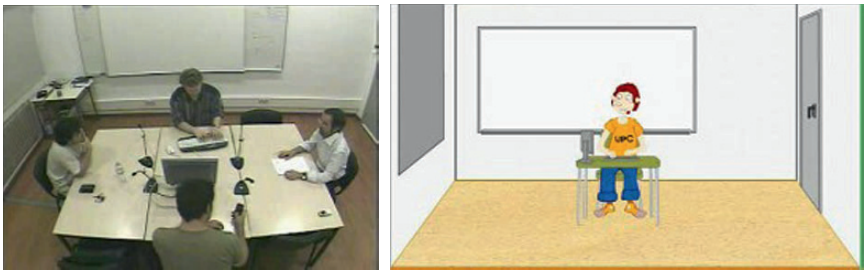
Fig. 7.1. Snapshot of the video showing the acoustic event detection and classification demo developed by FBK-irst.

A snapshot of the video appears in Fig. 7.1. The window on the left shows the position of the detected and classified event in the room map. The result of the classification process is represented through an icon placed in its estimated location in the room map and a text box on the bottom. The events in the room are shot by a camera whose output is reported in the window at the lower right corner. At the right upper corner are two windows showing the GCF and the oriented GCF used by the sound source localizer to estimate the location and directivity pattern of the acoustic event.

[1] The FBK-irst AED demo is available at http://www.youtube.com/watch?v=5_ZgrGL3CnU.

7.3.2 TALP-UPC Acoustic Event Detection and Classification Demo

The AED SVM-based system, written in the C++ programming language, is part of the smartAudio++ software package developed at UPC, which includes other audio technology components (such as speech activity detection and speaker identification) for the purpose of real-time online activity detection and observation in the smart room environment. That AED system implemented in the smart room has been used in demos about technology services developed in CHIL. Also, a specific GUI-based demo has been built that shows the detected isolated events and the positions of their sources in the room. The positions are obtained from the acoustic source localization (ASL) system developed at the TALP center [3].



(a) (b)

Fig. 7.2. The two screens of the GUI: (a) real-time video, and (b) the graphical representation of the AED and ASL functionalities ("keyboard typing" is being produced).

The video of the demo has been recorded while the demo was running in UPC's smart room[2]. It shows the output of the GUI during a session lasting about two minutes, where four people in the room speak, are silent, or make one of the 12 meeting-room sounds defined in CHIL and a few others. There are two screens in the GUI output, as shown in Fig. 7.2. One corresponds to the real video captured from one of the cameras installed in UPC's smart room, and the other is a graphical representation of the output of the two technologies. 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, thus showing the right event in a wrong place. The AED technology includes an "unknown" output, symbolized with "?". There are two different situations where the "unknown" label may appear. First and most frequently, it appears when the AED algorithm does not have enough confidence to assign a detected nonsilent event to one of the above-mentioned 12 classes. Second, the "unknown" label is produced when an out-of-list (never-seen-before) AE is detected.

[2] The UPC-TALP AED demo is available at http://www.youtube.com/watch?v=UBSxBd_HYeI.

7.4 Conclusions and Remaining Challenges

The work presented in this chapter has focused on several pioneering evaluation tasks concerning the detection and classification of acoustic events that may happen in a lecture/meeting-room environment. In this context, two different tasks have been evaluated: acoustic event classification and acoustic event detection. Two kinds of databases have been used: two databases of isolated acoustic events, and a database of interactive seminars containing a significant number of acoustic events of interest.

The evolution of task definition, the metrics, and the set of acoustic events have been presented and discussed. The detection and classification systems that submitted results to the evaluation campaigns have been succinctly described, and their results have been reported. In the classification task, the error rates shown in the last performed AEC evaluation (2006) were low enough to convince us to turn to the AED task, where the results obtained with acoustic events in isolated conditions have been satisfactory. However, the performances shown by the AED systems in meeting-room spontaneous conditions indicate that there is still much room for improvement.

References

1. G. B. D. Wang. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press, 2006.
2. R. Malkin, D. Macho, A. Temko, and C. Nadeu. First evaluation of acoustic event classification systems in the CHIL project. In *Joint Workshop on Hands-Free Speech Communication and Microphone Array, HSCMA'05*, March 2005.
3. C. Segura, A. Abad, C. Nadeu, and J. Hernando. Multispeaker localization and tracking in intelligent environments. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 82–90, Baltimore, MD, May 8-11 2007.
4. R. Stiefelhagen, R. Bowers, and J. Fiscus, editors. *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*. LNCS 4625. Springer, Baltimore, MD, May 8-11 2007.
5. R. Stiefelhagen and J. Garofolo, editors. *Multimodal Technologies for Perception of Humans, First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR'06*. LNCS 4122. Springer, Southampton, UK, Apr. 6-7 2006.
6. A. Temko. *Acoustic Event Detection and Classification*. PhD thesis, Universitat Politècnica de Catalunya, Barcelona, 2007.
7. A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo. Evaluation of acoustic event detection and classification systems. In *Multimodal Technologies for Perception of Humans. First International Evaluation Workshop on Classification of Events, Activities and Relationships CLEAR 2006*, LNCS 4122, pages 311–322. Springer-Verlag, Southampton, UK, Apr. 6-7 2006.