Acoustic Event Detection: SVM-based System and Evaluation Setup in CLEAR'07

Andrey Temko, Climent Nadeu, Joan-Isaac Biel

TALP Research Center, Universitat Politècnica de Catalunya (UPC),
Campus Nord, Ed. D5, Jordi Girona 1-3, 08034 Barcelona, Spain
{temko, climent, albiel}@talp.upc.edu

Abstract. In this paper, the Acoustic Event Detection (AED) system developed at the UPC is described, and its results in the CLEAR evaluations carried out in March 2007 are reported. The system uses a set of features composed of frequency-filtered band energies and perceptual features, and it is based on SVM classifiers and multi-microphone decision fusion. Also, the current evaluation setup and, in particular, the two new metrics used in this evaluation are presented.

1 Introduction

The detection of the acoustic events (AE) that are naturally produced in a meeting room may help to describe the human and social activity that takes place in it. 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.

After the Acoustic Event Detection (AED) evaluation within the CLEAR evaluation campaign 2006 [1] organized by the CHIL project [2], several modifications have been introduced into the task for the CLEAR evaluation campaign 2007. The old metric has been substituted by two new metrics: Accuracy and Error Rate, which are based, respectively, on precision/recall and on a temporal measure of detection error. Additionally, AED is performed only in seminar conditions, where the AEs are often overlapped with speech and/or other AEs. The definition of the classes of AEs is kept.

In this paper, after presenting the current evaluation setup and, in particular, the two new metrics used in this evaluation, we describe the AED system developed at the UPC and submitted to the CLEAR evaluations carried out in March 2007 along with its results.

The paper is organized as follows. In Section 2 the evaluation setup is presented. Specifically, the definition of the task is given in Subsection 2.1. Subsection 2.2 describes the databases assigned to development and testing. Metrics are given in Subsection 2.3, and Subsection 2.4 states the main evaluation conditions. The detailed description of the proposed system is given in Section 3. The results obtained by the

detection system in the CLEAR evaluations are shown and discussed in Section 4. Conclusions are presented in Section 5.

2 Evaluation setup

2.1 Acoustic Event classes

The AED evaluation will use the same 12 semantic classes, i.e. types of AEs, used in the past evaluations CLEAR 2006 [1]. The semantic classes with the corresponding annotation label are shown in black in the first column of Table 1. Apart from the 12 evaluated classes, there are 3 other possible events shown in grey in Table 1 which are not evaluated.

Table 1. Number of occurrences per acoustic event class for the development and test data

Event Type		Number of Occurrences			
		Development			Test
		UPC iso	ITC iso	Seminars	Seminars
Door knock	[kn]	50	47	82	153
Door open/slam	[ds]	120	100	73	76
Steps	[st]	73	50	72	498
Chair moving	[cm]	76	47	238	226
Spoon/cup jingle	[cl]	64	48	28	28
Paper work	[pw]	84	48	130	88
Key jingle	[kj]	65	48	22	32
Keyboard typing	[kt]	66	48	72	105
Phone ring	[pr]	116	89	21	25
Applause	[ap]	60	12	8	13
Cough	[co]	65	48	54	36
Laugh	[la]	64	48	37	154
Unknown	[un]	126	-	301	559
Speech	[sp]		-	1224	1239
Silence		Not annotated explicitly			

2.2 Databases

The database used in the CLEAR evaluation campaign 2007 consists of 25 interactive seminars of approximately 30 min long each that have been recorded by AIT, ITC, IBM, UKA, and UPC in their smart-rooms.

Five interactive seminars (one from each site) have been assigned for system development. Along with the seminar recordings, the databases of isolated AEs recorded at UPC [3] and ITC [4] have been used for development.

The development database details in terms of the number of occurrences per AE class are shown in Table 1. In total, development data consists of 7495 seconds, where 16% of total time is AEs, 13% is silence, and 81% is "Speech" and "Unknown" classes.

The remaining 20 interactive seminars have been conditionally decomposed into 5 types of acoustic scenes: "beginning", "meeting", "coffee break", "question/answers", and "end". After observing the "richness" of each acoustic scene type in terms of AEs, 20 5-minute segments have been extracted by ELDA maximizing the AE time and number of occurrences per AE class. The details of the testing database are given in Table 1. In total, the test data consist of 6001 seconds, where 36% are AE time, 11% are silence, and 78% are "Speech" and "Unknown" classes. Noticeably, during about 64% of time, the AEs are overlapped with "Speech" and during 3% they are overlapped with other AEs. In terms of AE occurrences, more than 65% of the existing 1434 AEs are partially or completely overlapped with "Speech" and/or other AEs.

2.3 Metrics

Two metrics have been developed at the UPC, with the agreement of the other participating partners which are involved in CHIL: an F-score measure of detection accuracy (which combines recall and precision), and an error rate measure that focuses more on the accuracy of the endpoints of each detected AE. They have been used separately in the evaluations, and will be called, respectively, AED-ACC and AED-ER.

AED-ACC

The aim of this metric is to score 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 to detect their instances. It is oriented to applications like real-time services for smart-rooms, audio-based surveillance, etc. AED-ACC is defined as the F-score (the harmonic mean between Precision and Recall):

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

where

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

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

and β is a weighting factor that balances Precision and Recall. In this evaluation the factor β has been set to 1. A *system output AE* is considered *correct* or *correctly produced* either if there exist at least one reference AE whose temporal centre is situated between the timestamps of the system output AE and the labels of the system output

AE and the reference AE are the same, or if the temporal centre of the system output AE lies between the timestamps of at least one reference AE and the labels of the system output AE and the reference AE are the same. A *reference AE* is considered *correctly detected* either if there exist at least one system output AE whose temporal centre is situated between the timestamps of the reference AE and the labels of the system output AE and the reference AE are the same, or if the temporal centre of the reference AE lies between the timestamps of at least one system output AE and the labels of the system output AE and the reference AE are the same.

AED-ER

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 [5] has been adapted to the task of AED. The audio data is 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 overlapping, in which a system output AE is not attributed correctly to a reference AE, in the following way:

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

where, for each segment *seg*:

dur(seg): duration of *seg*

N_{REF} *(seg)*: number of reference AEs in *seg*

N_{SYS} *(seg)*: number of system output AEs in *seg*

$N_{correct}$ *(seg)*: number of reference AEs in *seg* which correspond to system output AEs in *seg*

Notice that an overlapping region may contribute with several errors. Also, "Silence" is not explicitly transcribed, but is counted in the context of this metric as an AE.

The numerator of the AED-ER expression includes the substitution time, that corresponds to the wrong AE detection, the deletion time (missed AEs), and the insertion time (AE false alarms).

Only the 12 above-mentioned evaluated classes can cause errors. For example, if the reference label is "Speech" and the system output is "Unknown", there is no error; however if the system output is one of the 12 classes, it will be counted as an error (insertion). Similarly, if the reference is one of the 12 classes and the system output is "Speech", it will be also counted as an error (deletion).

2.4 Evaluation Scenario

In order to have systems comparable across sites, a set of evaluation conditions were defined [6]:
- The evaluated system must be applied to the **whole** CLEAR 2007 test DB.
- Only **primary** systems are submitted to **compete**.
- The evaluated systems must use **only audio** signals, though they can use **any** number of **microphones**.

3 Acoustic Event Detection System

The general scheme of the proposed system for AED is shown in Figure 1. Firstly, on the data preprocessing step, the signals are normalized based on the histograms of the signal energy. Then, a set of frame-level features is extracted from each frame of 30ms and a set of statistical parameters is computed over the frames in a 1-second window. The resulting vectors of statistical parameters are fed to the SVM classifier associated to the specific microphone. A single-microphone post-processing is applied to eliminate uncertain decisions. At the end, the results of 4 microphones are fused to obtain a final decision.

Our system, written in 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 activity detection and observation in the smart-room environment. That AED system implemented in the smart-room has been used in the demos about technology services developed in CHIL. Also, a specific GUI-based demo has been built which shows the detected isolated events and their positions in the room. The positions are obtained from the acoustic source localization system developed also in our lab [11]. A video showing that demo is being currently recorded and will shortly be made publicly available in the CHIL webpage.

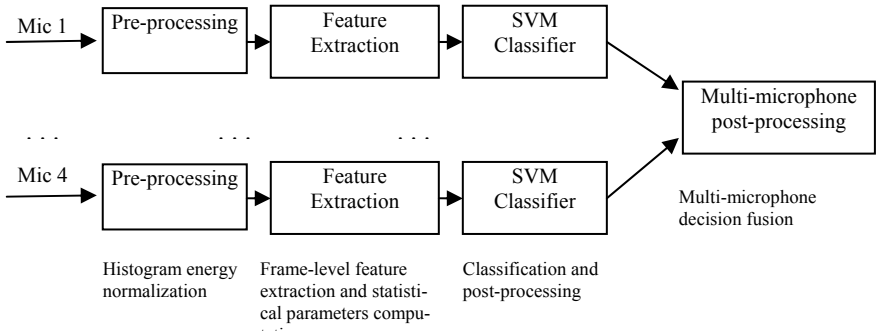

Figure 1. The block-scheme of the developed AED system

3.1 Histogram-based energy normalization

As it was mentioned in Section 2.2 the evaluation database has been recorded in 5 different rooms. Due to this fact, the energy level of audio signals varies from one

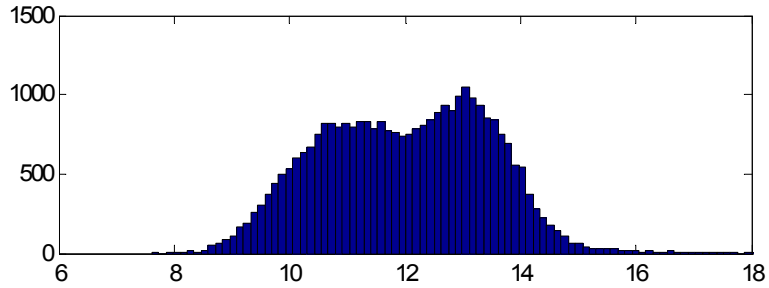

Figure 2. Frame log-energy histograms calculated over the whole seminar signal

audio file to another. In this work as a preprocessing step we decided to perform energy normalization of all audio files to a predefined level. Because the energy level of a given AE depends both on its type, the manner it is produced, and the position of the person who produces it, the energy normalization is based on the energy level of silence. For this the histogram of the audio signal log-energy calculated each 30ms with 10ms shift has been plotted. The results for one development seminar are shown in Figure 2. The lower-energy hump corresponds to the silence energy level. A 2-Gaussians GMM has been trained on the energy values and the lowest mean has been taken as the estimation of the silence energy. In Figure 2, the estimated silence level corresponds to the point 10.41 whereas the true value of silence energy level, calculated on the annotated silence segments, is 10.43. The normalizing coefficient is then calculated as $coef = \sqrt{\exp(9)/\exp(a)}$, where a is the estimated silence level and 9 is the predefined final silence energy level. The exponential is used to come from the log scale back to the initial signal amplitude scale. Then, the development seminar signal is multiplied by *coef*.

3.2 Feature extraction

The sound signal is down-sampled to 16 kHz, and framed (frame length/shift is 30/10ms, a Hamming window is used). For each frame, a set of spectral parameters has been extracted. It consists of the concatenation of two types of parameters [7]: 1) 16 Frequency-Filtered (FF) log filter-bank energies along with the first and the second time derivatives, and 2) a set of the following parameters: zero-crossing rate, short time energy, 4 sub-band energies, spectral flux, calculated for each of the defined sub-bands, spectral centroid, and spectral bandwidth. In total, a vector of 60 components is built to represent each frame. The mean and the standard deviation parameters have been computed over all frames in a 1-second window with a 200ms shift, thus forming one vector of 120 elements.

3.3 One-microphone SVM system

For AED, SVM classifiers [8] have been implemented. They have been trained using the isolated AEs from the two databases of isolated acoustic events mentioned in Section 2.2, along with segments from the development data seminars that include both isolated AEs and AEs overlapped with speech. The segments that contain the overlapping of two or more AEs with or without speech are not used. In both training and testing processes, a vector of 120 statistical parameters has been computed from each 1-second window. The 1 vs. 1 multiclass strategy has been chosen to classify among 14 classes that include "Speech", "Unknown", and the 12 evaluated classes of AEs. Besides, "Silence" vs. "Non-silence" SVM classifier has been trained where "Non-silence" class includes all 14 classes. In that case, in order to decrease the number of training vectors and make training feasible, the dataset reduction technique described in [9] has been applied.

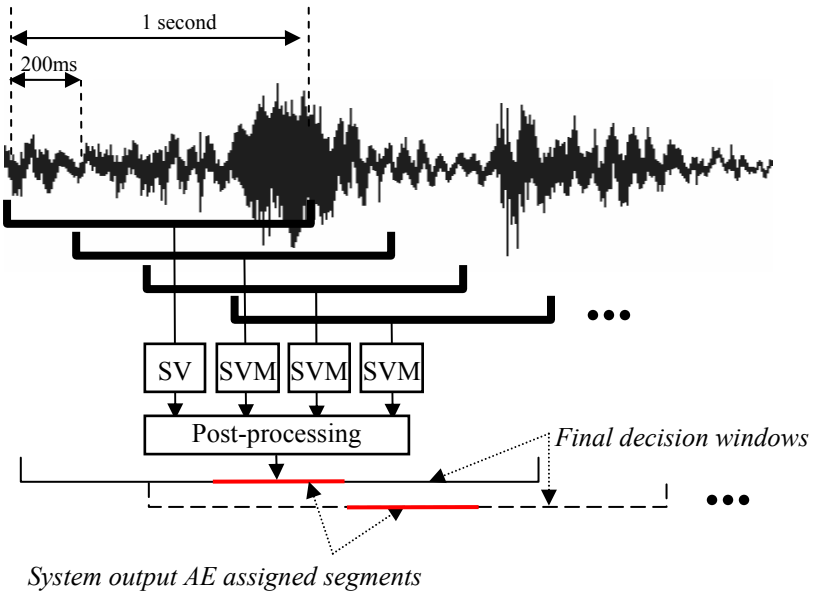

Figure 3. One microphone AED system

The testing stage is shown in Figure 3. An input vector of statistical components computed over the frames from a 1-second window is firstly fed to the "Silence" vs. "Non-silence" classifier and if the decision is "Non-silence", the vector is further fed to a SVM multiclass (14 classes) classifier based on the DAG testing scheme [10]. The most frequent event (the "winner") is taken from the final decision window of 4 decisions that corresponds to the time interval of 1.6 seconds. If the number of votes of the "winner" does not exceed the threshold the event is marked as "Unknown". The threshold has been set in order that the winner has to get at least 3 votes. The final decision window is shifted by 2 decisions, i.e. 400ms. Consequently, the tempo-

ral resolution of the produced system output AEs is 400ms, and the corresponding AE label is assigned to the central 400ms of the 1.6-second window.

For instance, for the first window of 4 decisions that corresponds to the time interval from 0 to 1.6s, the starting and ending timestamps of the system output AE will be 0.6 and 1s.

3.4 Multi-microphone processing

The database used in the evaluation has been recorded with a set of microphones. Depending on the site, the following audio equipment has been used: one or two Mark III (array of 64 microphones), 3-7 T-shape clusters (4 mics per cluster), and several tabletop and omni directional microphones. To construct a multi-microphone AED system it has been decided to choose one microphone from each wall of the room and train a SVM classifier for each wall microphone. Due to the different configuration of the rooms where the development and testing data have been recorded and due to different numbering of the microphones, a mapping of the microphones across the sites has been performed. The Mark III microphone array has been chosen as the fixed point. For the remaining walls the T-shape cluster microphones have been chosen. An example of choice of the cluster microphones for the UPC smart-room is shown in Figure 4. The following microphone numbers have been chosen 1-5-9, 6-1-25, 1-5-9, 1-5-9 for the AIT/ITC/UKA/UPC smart-rooms, respectively. For instance, one SVM has been trained on audio signals from microphones 1, 6, 1, 1 taken from AIT/ITC/UKA/UPC, respectively. For the Mark III array the 3rd microphone has been chosen across all sites.

For multi-microphone decision fusion, the voting scheme has been used. The AE label with the largest number of votes is sent to the system output. In case of draw the event is chosen randomly.

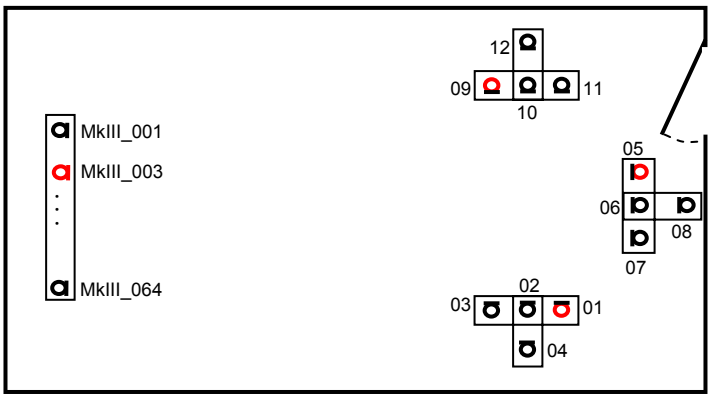

Figure 4. The choice of the microphones for the UPC smart-room

4 Results and discussion

The results obtained with the primary system submitted to the evaluation are shown in Table 2. Along with the main metrics, accuracy and error rate, the intermediate values are also given. They are precision and recall for accuracy, and DEL (deletions), INS (insertions), and SUB (substitutions) for error rate. A contrast system has been also submitted, showing little worse results than the primary system: ACC=23, ER=141.57. The difference between the primary and contrast system is that for multi-microphone fusion the former uses voting among the "winners" of the one-microphone systems while the contrast system performs voting adding up the confidences of the "winners" calculated as the number of times the "winner" is found in the 4-decision window.

Table 2. Official results obtained by the submitted AED primary system

Accuracy (%) (Precision / Recall)	Error Rate (%) (DEL/INS/SUB)
23.0 (19 / 29)	136.69 (50.3 / 57.1 / 29.3)

Table 3 shows the results of each one-microphone SVM system before applying the voting decision. Actually, the final results of the multi-microphone system shown in Table 2 are worse that the results of the one-microphone SVM system obtained on the 3rd microphones of MarkIII array (Mic4). This fact may indicate that simple fusion methods, i.e. voting, do not work properly when the scores of the various systems differ significantly.

Table 3. The results obtained with each one-microphone SVM system before applying voting

	Mic1	Mic2	Mic3	Mic4
Accuracy (%) (Precision / Recall)	20.5 (17/27)	22.6 (19/28)	19.9 (15/29)	**26.8** **(34/22)**
Error Rate (%) (DEL/INS/SUB)	145 (51/64/30)	136 (54/55/27)	155 (46/74/34)	**98** **(69/13/16)**

The individual class accuracies are shown in Table 4. Interestingly enough, we have observed that the low accuracy and high error rate are mostly attributable to the bad recognition of the class "steps", which occurs more than 40% of all AE time.

Besides, more than 76% of all error time occurs in the segments where AEs are overlapped with speech and/or other AEs. If the overlapped segments were not scored, the error rate of the primary submitted system would be 32.33%.

Table 4. Accuracy scores for each class obtained with the primary system

ap = 0.81	cl = 0.29	cm = 0.22	co = 0.19
ds = 0.42	kj = 0.18	kn = 0.05	kt = 0.08
la = 0.38	pr = 0.28	pw = 0.12	st = 0.16

5 Conclusions

The presented work focuses on the CLEAR evaluation task concerning the detection of acoustic events that may happen in a lecture/meeting room environment. The evaluation has been performed on the database of interactive seminars that have been recorded in different smart-rooms and contain a significant number of acoustic events of interest. Two different metrics have been proposed and implemented. One is based on the precision and recall of the detection of the AEs as semantic instances, and the other is based on a more time-based error. Although the proposed system, which was the only submission not using HMM, ranked among the best, there is still a big room for improvement. Future work will be devoted to search a better way to deal with overlapping sounds, and to improve the algorithms of multi-microphone fusion. Multimodal AED is another approach from which a performance improvement can be expected.

Acknowledgements

This work has been partially sponsored by the EC-funded project CHIL (IST-2002-506909) and the Spanish Government-funded project ACESCA (TIN2005-08852). 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.

References

1. A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, M. Omologo, “CLEAR Evaluation of Acoustic Event Detection and Classification systems”, *CLEAR'06 Evaluation Campaign and Workshop, Southampton*, LNCS, vol.4122, pp.311-322, Springer, January 2007
2. CHIL - COMPUTERS IN THE HUMAN INTERACTION LOOP, http://chil.server.de/
3. A. Temko, D. Macho, C. Nadeu, C. Segura, “UPC-TALP Database of Isolated Acoustic Events”, *Internal UPC report*, 2005
4. C. Zieger, M. Omologo, “Acoustic Event Detection - ITC-irst AED database”, *Internal ITC report*, 2005
5. “Spring 2007 (RT-07) Rich Transcription Meeting Recognition Evaluation Plan”, NIST, December, 2007
6. A. Temko, C. Nadeu, “AED Evaluation plan for CLEAR 2007”, Feb. 2007
7. A. Temko, C. Nadeu, “Classification of Acoustic Events using SVM-based Clustering Schemes”, Pattern Recognition, volume 39, issue 4 , pp.682-694, Elsevier, April 2006
8. B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002
9. A. Temko, D. Macho, C. Nadeu, “Enhanced SVM Training for Robust Speech Activity Detection”, *IEEE ICASSP 2007*, Honolulu, Hawaii, USA, April 2007
10. J. Platt, N. Cristianini, J. Shawe-Taylor, “Large Margin DAGs for Multiclass Classification”, *Proc. Advances in Neural Information Processing Systems* 12, pp. 547-553, 2000
11. C. Segura, A. Abad, J. Hernando, C. Nadeu, “Multispeaker Localization and Tracking in Intelligent Environments”, *CLEAR'07 Evaluation Campaign and Workshop*, Baltimore MD, USA, May 2007