

# TUT Audio Event Detection System 2007

Toni Heittola and Anssi Klapuri

Tampere University of Technology, P.O.Box 553, 33101, Tampere, Finland  
{toni.heittola,klap}@tut.fi

**Abstract.** This paper describes a system used in acoustic event detection task of the CLEAR2007 evaluation. The objective of the task is to detect acoustic events (door slam, steps, paper wrapping etc.) using acoustic data from a multiple microphone set up in the meeting room environment. A system based on hidden Markov models and multi-channel audio data was implemented. Mel-Frequency Cepstral Coefficients are used to represent the power spectrum of the acoustic signal. Fully-connected three-state hidden Markov models are trained for 12 acoustic events and one-state models are trained for speech, silence, and unknown events. The system performed adequately compared to other participant in CLEAR2007 evaluation.

## 1 Introduction

In the meeting room environment, the identification of acoustic events may help to describe activity that takes place in the room. For example, door slams and chair movements can be used to detect when the meeting has begun or ended. The event activity information is essential for perceptually aware interfaces in smart meeting rooms. Computational auditory scene analysis may be used to detect and identify acoustic events [1]. Acoustic Event Detection (AED) and Classification aims to process the acoustic signal and convert it into a symbolic descriptions of the corresponding sound events present in the acoustic signal. Information produced by the AED system can be further used to increase the robustness of automatic speech recognition, for example.

Acoustic event detection task is closely related to the more general task of noise classification and recognition [2]. Audio-based context-awareness systems [3,4,5] classify acoustical environments based on the general acoustic information of individual acoustic events and these systems utilize many of the approaches used in AED systems.

In this paper, we present our system for the acoustic event detection task in CLEAR2007 evaluation [6]. The next section discusses the evaluation task and the data and the metrics that are used. Then a description about the proposed system is given. In Section 3.4 results of the evaluation are given. Section 4 conclude the discussion.

## 2 System Description

The proposed system is based on modeling sound events with continuous-density hidden Markov models (HMMs). Mel-Frequency Cepstral Coefficient are used to represent the power spectrum. Observation probabilities are calculated separately for one channel from each T-shaped microphone array. After that, all the observation probabilities are combined (see Figure 1) and the optimal path through all models (see Figure 2) is decoded.

### 2.1 Features

Mel-Frequency Cepstral Coefficients (*MFCCs*) and their time-derivatives are the most widely-used features in speech recognition [7], and recently these have been successfully used in audio-based context recognition [3,4,5] and noise classification tasks [1,2], too. MFCCs are short-term feature used to represent the coarse shape of the power spectrum and provide a good discriminative performance with reasonable noise robustness. We used 10 MFCCs calculated from the outputs of a 40-channel filterbank. In addition to the static coefficients, their differentials ( $\Delta MFCC$ ) are used to describe the dynamic properties of the cepstrum. Signal's log-energy is also used in the feature vector. Before the feature extraction, the time-domain signals are normalized to have zero mean and unity variance over the training data. Features are extracted in 23 ms frames with 50% overlap. In order to normalize frequency responses of microphones (channel effect) used in the recordings, all the features are mean and variance normalized using global estimates measured over the all data available from a particular audio channel.

### 2.2 Classification

Hidden Markov models (HMMs) are used to characterize a time-varying series of observations. Fully-connected three-state HMMs having four Gaussians per state are trained for each of the 12 acoustic events. One state HMMs with 32 Gaussians are trained for speech, silence, and unknown events. Training segments for event models are selected from the development set by preferring event segments which are not overlapping with speech. Amount of training segments used per event are shown in Table 1. All the available audio channels are used for the training. HMMs are trained with the standard Baum-Welch training procedure.

In the classification stage, observation probabilities are calculated for each microphone channel. These probabilities are then combined by presuming microphone channels independent, e.g. probabilities are multiplied (see Figure 1). Finally the optimal path through all models is decoded with Viterbi algorithm. All transitions between models are set equiprobable. An overview of model structure is shown in Figure 2. The final event sequence is smoothed with one-second sliding window.

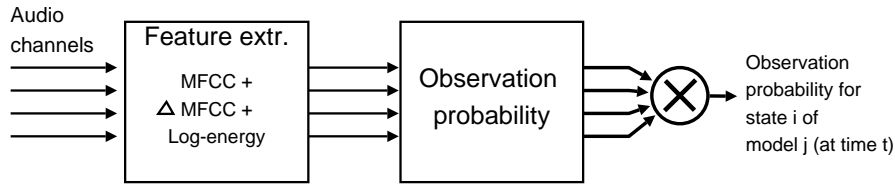


Fig. 1. Calculation of observation probability for multichannel audio data.

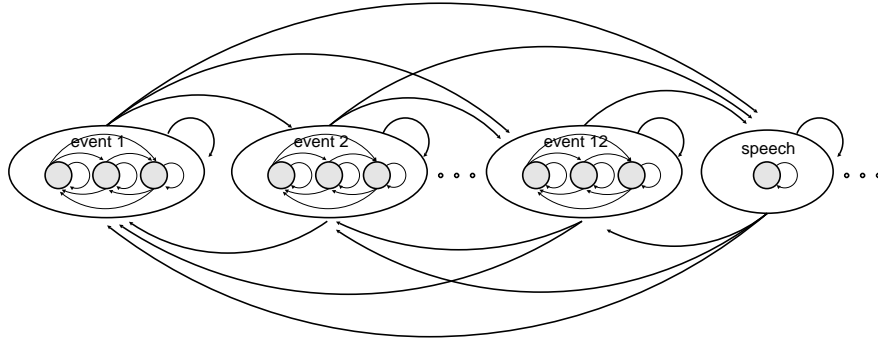


Fig. 2. Model overview.

### 3 Evaluation

#### 3.1 Task

The objective of the task is to detect and recognize acoustic events using only the acoustic data from the multiple microphones placed in the meeting room environment. The evaluation will use the 12 semantic event classes shown in Table 1.

#### 3.2 Data

A description about the database used in system development and performance evaluation is given here. For further details, refer to the CLEAR2007 evaluation plan [6]. The data is divided into development (Dev) and evaluation (Eval) sets. Segments for these sets are extracted from five interactive seminar databases (one seminar per database are selected for the Dev and 8 per database for Eval):

- Politècnica de Catalunya (UPC) Interactive Seminar Database 2006
- Institutino di Cultura (ITC) Interactive Seminar Database 2006
- Athens Information Technology (AIT) Interactive Seminar Database 2006
- University of Karlsruhe (UKA) Interactive Seminar Database 2006
- IBM Interactive Seminar Database 2006

The data consists of about 30 minute audio recordings made in the meeting room environment while a presentation is given to a group of 3-5 attendees. There is typical meeting room activity present: discussion among attendees and presenter, people entering and leaving the room, people having coffee breaks, chairs moving, paper wrapping, etc. The acoustic events in the recordings may have temporal overlapping with speech and/or other acoustic events. The recordings are annotated by hand with a time resolution of 150 ms. Annotated event classes and the number of segments per event are shown in Table 1. The recordings were made at 44.1kHz sampling rate and at 24 bit resolution. The number of microphone arrays available depends on the configuration of the site, see Table 2.

**Table 1.** Number of events in the seminar database.

Event	Total count	Selected for training
Knock (door,table)	77	36
Door slam (open and close)	71	33
Steps	89	25
Chair moving	263	39
Spoon clings (cup jingle)	29	9
Paper wrapping	163	9
Key jingle	21	4
Keyboard typing	73	8
Phone ringing/music	27	2
Applause	8	2
Cough	54	17
Laugh	43	5
Speech	1230	
Unknown	317	

**Table 2.** Information about the data used in the evaluation is presented.

Site	Amount of data [minutes]		
	T-arrays	Dev set	Eval set
UPC	3	23	41
ITC	7	31	41
AIT	3	31	41
UKA	4	44	41
IBM	4	37	40

### 3.3 Metrics

Two metrics will be used in evaluations. An F-score is used to measure detection accuracy (AED-ACC) and error rate (AED-ER) is used to measure the accuracy

of the endpoints of detected acoustic events. Only the 12 evaluated classes can cause errors in the evaluations, e.g. speech events are allowed to be confused with unknown events and vice versa.

**Accuracy** Accuracy (AED-ACC) metric is used to evaluate detection of all relevant acoustic events. With this metric, a good temporal coincidence of the reference and system output is not important. The metric is oriented to applications audio-based services in smart meeting rooms and surveillance systems. AED-ACC is similar to F-score and it is defined as the harmonic mean between precision and recall:

$$AED - ACC = \frac{(1 + \beta^2) * precision * recall}{\beta^2 * precision + recall}, \quad (1)$$

where

$$precision = \frac{\text{number of correct system output events}}{\text{number of all system output events}},$$

$$recall = \frac{\text{number of correctly detected reference events}}{\text{number of all reference events}}$$

and  $\beta$  is a weighting factor (in the evaluations set to one) that balances precision and recall. An event detected by the system is considered correct if there exist at least one matching reference event having its temporal center within timestamps of detected event. Detection is also considered correct if the temporal center of detected event is within timestamps of at least one matching reference event. A reference event is considered correctly detected if the temporal center of reference event is within at least one matching system output or if there exist at least one matching system output having its temporal center within timestamps of the reference event.

**Error rate** Error rate (AED-ER) metric is used to evaluate temporal resolution. This is done by scoring the accuracy of the endpoints of the detected events. The NIST metric for Speaker Diarization [8] was adapted to the task of acoustic event detection. First an one-to-one mapping of reference events to system output events will be computed. The measure of mapping optimality will be the aggregation, over all reference events, of the time length that is jointly attributed to both the reference events and the corresponding system output events to which that reference events are mapped. This metric is oriented to applications involving content-based audio indexing and segmentation.

The audio signal is divided into contiguous segments, whose borders coincide with starts or stops of a reference or a system output event. For a given segment, the number of current reference events and the number of system output events do not change. The AED-ER is defined as the fraction of the time that is not attributed correctly to an acoustic event:

$$AED - ER = \frac{\sum_{all\ seg} \{dur(seg) * (max(N_{REF}, N_{SYS}) - N_{correct}(seg))\}}{\sum_{all\ seg} \{dur(seg) * N_{REF}(seg)\}} \quad (2)$$

where  $dur(seg)$  is the duration of the  $seg$ ,  $N_{REF}(seg)$  is the number of reference events in  $seg$ ,  $N_{SYS}(seg)$  is the number of system output events in  $seg$  and  $N_{correct}(seg)$  is the number of reference events in  $seg$  which have corresponding mapped system output events in  $seg$ . A wrong event detection, the deletion time (missed events), and the insertion time (false alarms) corresponds to the substitution time included in the numerator.

### 3.4 Results

The performance was evaluated based on the the AED-ACC and AED-ER metrics. The proposed system obtained accuracy score 14.7 with the AED-ACC metrics, precision was 19.2 (164/853) and recall 11.9 (170/1434). Overall detection error with AED-ER metric was 139.06 percent of scored event time, see detailed results in Table 3. All the simulations were made with Matlab running on the Intel Pentium 4 CPU 3.2GHz, and the total computational time required for the performance test was 8906 seconds.

**Table 3.** Detailed results with the AED-ER metric.

Scored acoustic event time	2167.75 secs	37.4 % scored time
Missed acoustic event time	1189.68 secs	54.9 % scored acoustic event time
False alarm acoustic event time	1090.95 secs	50.3 % scored acoustic event time
Event substitution time	733.86 secs	33.9 % scored acoustic event time
Overall detection error	139.06 %	

In Table 4 is shown results for the all CLEAR07/AED task participants. Although our system utilized rather traditional approach for the signal classification, in comparison to the other participants it still performed rather adequately.

**Table 4.** AED task results for all the CLEAR07/AED participants.

Participant	AED-ACC	AED-ER	comment
UIUC	36.3	99.49	
UPC	23.0	136.7	
STI2R	22.9	170.49	
<b>TUT</b>	<b>14.7</b>	<b>139.06</b>	
ITC/FBK	23.4	109.07	site-dependent
AIT	4.4	203.11	site-dependent

## 4 Conclusions

We presented a system used for the acoustic event detection task in CLEAR2007 evaluation. System utilized widely used classification system scheme, power spectrum based features which were classified with HMM classifier. Our system was found to give acceptable performance compared to other participant in CLEAR2007 evaluation.

In general, the acoustic event detection is a difficult task due to overlapping events, noise and acoustic variation. Multi-channel audio information has to be more utilized in the detection in order to better performance. In addition to this, more robust feature extraction techniques or sound separation is also needed.

## References

1. Temko, A., Nadeu, C.: Classification of acoustic events using SVM-based clustering schemes. *Pattern Recogn.* **39**(4) (2006) 682–694
2. Gaunard, P., Mubikangiey, C., Couvreur, C., Fontaine, V.: Automatic Classification of Environmental Noise Events by Hidden Markov Models. *Applied Acoustics* **54**(3) (1998) 187–206
3. Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., T., S.: Computational Auditory Scene Recognition, *International Conference on Acoustic, Speech and Signal Processing*
4. Eronen, A., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., Huopaniemi, J.: Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing* **14**(1) (January 2006)
5. Ma, L., Milner, B., Smith, D.: Acoustic Environment Classification. *ACM Trans. Speech Lang. Process.* **3**(2) (2006) 1–22
6. CLEAR: AED Evaluation Plan 2007. [http://isl.ira.uka.de/clear07/?download=CLEAR\\_2007\\_AED\\_EvaluationPlan.pdf](http://isl.ira.uka.de/clear07/?download=CLEAR_2007_AED_EvaluationPlan.pdf)
7. Rabiner, L., Juang, B.H.: *Fundamentals of Speech Recognition*. PTR Prentice-Hall Inc., New Jersey (1993)
8. NIST: Spring 2005 (RT-05S) Rich Transcription Meeting Recognition Evaluation Plan. <http://nist.gov/speech/tests/rt/rt2005/spring/rt05s-meeting-eval-plan-V%1.pdf>