

The Acoustic Event Detector of AIT

C. Boukis and L.C. Polymenakos

Athens Information Technology, Greece

Abstract. In this paper the acoustic event detection and classification system that has been developed at Athens Information Technology is presented. This system relies on the use of several Hidden Markov Models arranged in a hierarchical manner in order to provide more accurate detections. The audio streams are split into overlapping frames from which the necessary for training and testing features are obtained. A post processing scheme has also been developed in order to smooth the raw detections. The results that were obtained from the application of this system on the testing data of the CLEAR evaluation, obtained from five different sites are presented and the performance of this system is discussed.

1 Introduction

Acoustic event detection (AED) is a tedious task whose objective is the detection of specific acoustic events within one or more synchronised audio streams. These audio streams can be processed either in an online fashion (while they are captured from the microphones) or offline (after being captured and stored). The list of event that we look for is pre-defined and *a-priori* data for every event are required.

The detection of acoustic events within audio streams and their subsequent classification is a relatively new research area. Existing systems rely on the use of classification techniques. Typical examples is the use of Hidden Markov Models (HMM) [1] or Support Vector Machines (SVM) [2]. These approaches require training data for the optimisation of the parameters of their classification systems . These training data are collected from the same environment in which the AED system will be evaluated, in order to be as coherent as possible with the testing data, covering several noise conditions. Many approaches use a silence detector that prevents the classification of silence intervals as events and thus increases the overall performance of the system [3].

AED can be considered as a generalisation of voice activity detection (VAD) [5, 6], where the objective is the detection of speech intervals within audio signals. VAD can be performed either with unsupervised methods that employ statistical criteria like the Likelihood Ratio Test (LRT) [7], or with supervised methods like Hidden Markov Models (HMM) and Linear Discriminant Analysis (LDA) [4]. Since AED attempt to identify several acoustic events, only supervised methods are used since only these can distinguish adequately various acoustic events. Similarities can be found between AED and face or object recognition within

images. In both cases the detection of a specific pattern within a data vector is desired. In object (or face) recognition though it is assumed that the entire object (or face) is contained entirely in an image; this is not the case in AED where an acoustic event might last from one to several frames.

In this paper the primary and the contrast AED systems that have been developed by Athens Information Technology are presented. These two systems rely on the use of HMMs for the identification of acoustic events. Their difference is that for the primary system employs a different set of HMM parameters for every site, which is derived by using only the data derived from this site for training, while the contrast system uses the same set of HMM parameters for every site, derived by applying a global training scheme using all the available data. The use of hierarchical structures for the improvement of the performance of the system is also proposed and post processing methods that aim at the smoothing of raw decisions are discussed.

This paper is organised as follows: In section 2 the pre-processing of the audio signals for both the development and the testing data is discussed. In section 3 the HMM-based decision-making process is presented. Post-processing of the results is shown in section 4, while the results are presented in section 5. Finally, section 6 concludes the paper.

2 Pre-Processing

The AED system uses audio signals obtained from one (or more) microphone, which are used either as development data for training or as hypothesis data for decision making. These audio signals are initially down-sampled (or interpolated) in order for all of them to have the same sampling rate, and subsequently they are split into frames. More analytically:

Framing: Detection using supervised methods, like HMMs or SVMs, requires the processing of data in frames. Every time a frame of available data is passed into the system and either the parameter estimates are improved if it is in training mode, or a decision is made if it is in decision-making mode. In our approach we have chosen the frame length to be 2048 samples. The overlapping between neighbouring frames was 75%, that is 1536 samples. These values correspond to 93 *msecs* duration and 70 *msecs* overlapping. These values were chosen so as many short time acoustics events like door knock, door slam, clapping etc to be included within a single frame.

Down-sampling: Since the number of samples that are included within a frame has been chosen to be fixed, all the audio signals should have the same sampling rate, so as frames to have the same duration. Therefore development and hypothesis data should have the same sampling rate. To this cause the hypothesis data are down-sampled by a factor of two, since their sampling rate (44100 Hz) is twice that of the development data.

Feature extraction: In classification applications with unsupervised techniques it is usually more convenient, more robust and effective to extract some features from the data frames and process them instead of the raw data

(time-domain audio samples in our case). For the sake of acoustic event detection the feature that have been chosen to operate with are the first 12 mel frequency cepstral coefficients along with the energy and their 1st and 2nd order derivatives. These together with the zero-crossing rate consist the 40-elements long feature vectors that have been used in the present AED implementation.

Notice that the down-sampling and the framing process might affect significantly the performance of the AED system.

Table 1. Notation of acoustic events

event	notation
door slam	ds
steps	st
chair moving	cm
spoon(cup jingle)	cj
applause	ap
laugh	la
key jingle	kj
cough	co
keyboard typing	kt
phone ringing/music	pr
knock (door, table)	kn
paper wrapping	pw
speech	sp
unknown	un

2.1 Development Data

The development data consist of the audio part of five seminars each one recorded at a different site (AIT, IBM, ITC, UKA, UPC) and two isolated events data bases provided by ITC-irst and UPC respectively. The seminar recordings were monophonic, their sampling rate was 22050 Hz and they were captured using the markIII microphone array, while the isolated events databases contained recordings from several microphones (hamerfall and markIII microphone array) at 44100 Hz [Ref]. Accompanying csv files were provided for each recording including time-stamps about the begin and the end of every event.

Using the time-stamps provided in the csv files the occurrences of every event in the seminar files were obtained. These occurrences were subsequently split into frames according to the procedure described earlier. For every frame a set of features was extracted, which were sub-sequently used for the training

of the parameters of the HMMs. Notice that the provided csv files contain information concerning the 12 acoustic events that we are looking for, along with information concerning speech occurrences and unknown sounds (Table 1). The seminar segments that are not annotated at all were treated as silence and were used for the training of the HMM that were developed for the modelling and the detection of silence intervals.

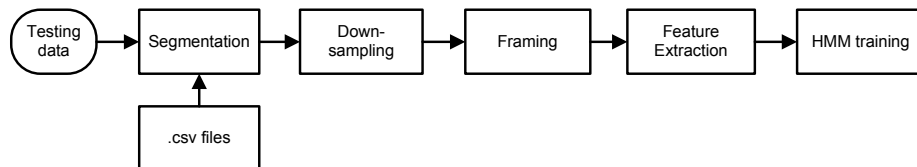


Fig. 1. Training of the HMMs for the task of AED

In order for the development and the testing data to be coherent only data obtained from the markII microphone array were used. Hence only the acoustic events of the UPC database were employed. Since The procedure that was followed for the extraction of feature vectors from the isolated events data bases was the following: the recordings from five different microphones of the array were averaged in order to enhance the waveforms of the acoustic events and to suppress the ambient noises. The microphones that were used were the 13th, the 23rd, the 33rd, the 43rd and the 53rd microphones. The obtained signal was subsequently down-sampled in order to reduce its sampling rate from 44100 Hz to 22050 Hz. Using the provided csv files the occurrences of every event were extracted and they were subsequently split into overlapping frames. Finally, for every frames a corresponding feature vector was computed. The training process is presented graphically in the block diagram of Fig. 1. Notice that the silence intervals of the isolated seminars were not used as training data, since they are not coherent with the silence intervals of the seminars which would be used for the evaluation of the system.

2.2 Hypothesis Data

The hypothesis data that were used for the evaluation of the developed system are 20 segments of seminars recorded in 4 different sites: AIT, ITC, UKA and UPC. The duration of each segment is approximately 5 *mins* (300 *secs*) and the sampling rate is 44100 Hz. For each seminar segment recordings captured from the 64 microphones of the mark III array and the hammerfall microphones were provided.

In order to produce decisions the signals of the microphones 13, 23, 33, 43 and 53 were averaged. This form of spatial averaging was applied in order to suppress the ambient noises and enhance the occurrences of the acoustic events. every seminar segment was initially down-sampled to 22050 Hz. The down-sampled signal

was subsequently split into frames of length 2048 samples. Neighbouring frames were overlapping by 75% (1536 samples). For every frame a feature vector was derived that was fed to the acoustic event detector. This process is graphically illustrated in Fig. 2

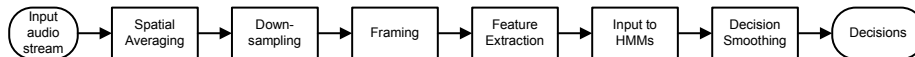


Fig. 2. Decision making process

3 Decision Making

The developed AED system relies on the use of HMMs. More specifically left-right (or Bakis) models were employed with continuous observation densities. Each model was characterised by its number of states N and the transition matrix A , which has the form

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & 0 \\ 0 & a_{22} & a_{23} & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{N-1N} \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \quad (1)$$

for left-right model of order N . The initial state probability of the first state is 1, while that of all the other states is zero, that is

$$\pi_i = \begin{cases} 0 & , \text{ when } i \neq j \\ 1 & , \text{ when } i = j \end{cases} \quad (2)$$

The continuous observation density of every state i is modelled with a single Gaussian with mean vector μ_i and covariance matrix Σ_i . For the training of the parameters of every HMM the Baum-Welch method was employed. The number of states of every model was deduced heuristically and these are presented in Table 2.

3.1 Initial System

Initially an HMM model for every event was developed. Two training policies were applied aim at the maximisation of the discrimination power of these models

- Partial training: For every site derive a separate set of HMMs by performing training only with the corresponding development data and the isolated

Table 2. Number of states of employed HMMs

Event	Number of states
ap	3
cl	3
cm	3
co	3
ds	3
kj	3
kn	3
kt	3
la	4
pr	4
pw	3
sl	3
sp	8
st	3
un	3

events databases. In this case the HMMs that are used for the detection of acoustic events within AIT’s seminars were trained using the development data of AIT and the isolated events databases of UPC. Similarly the HMM models used for detection of events in the IBM seminars were trained using the IBM development data together with the UPC isolated events databases and so on.

- Global training: In this scenario a unique set of HMMs was used for the detection of events in every site. Their parameters were obtained by performing training using all the available seminar development data along with the isolated events database of UPC.

3.2 Hierarchical Detection

The initially developed AED system was under-performing since it was continuously misclassifying silence intervals as event occurrences. To improve its performance detection process was split into two stages, and thus became a two-step process. In the first step two left-right HMMs were used to detect the existence of an event or silence. The silence model was of order 3 and it was trained using the silence intervals that are inherent in the seminar development data, while for the training of the event HMM, whose order was 8, all the event data were used. In this AED implementation the type of the event is detected in the second stage by using 14 HMM models, provided that the presence of an event has been deduced in the first layer of this structure. The orders of the employed HMMs are presented in Table 2. Partial and global training was applied

in this case two and it was found out that performing detection in stages results in improved performance. The system that was derived with partial training is the primary system that was used in the CLEAR evaluation, while the system that was derived with global training is the contrast system. This hierarchical structure is depicted in Fig. 3.

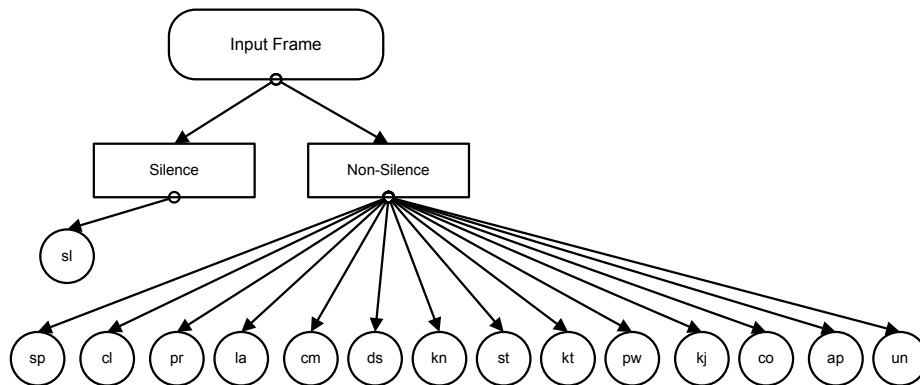


Fig. 3. A two-layer hierarchical structure for AED using HMMs.

Grouping acoustic events together, based on their characteristics can lead to multilayered hierarchical detection approaches that are expected to have improved performance compared to conventional methods. In this approach the existence of an event is examined in the first layer of the hierarchy. If the content of this frame is not silence it is examined whether this is speech or not. The HMM model that is used for the detection of speech is trained using all speech data, while all the other data are used for the training of the coefficients of the non-speech HMM. If the decision is non speech, it is examined whether the frame contains a periodic or a non-periodic event and so on. Every internal layer of decision has two HMMs while the lower layer of the hierarchy (the leaves of the tree) can have several HMMs. An example of a multilayer approach is shown in Fig. 4.

4 Post-processing

The raw decisions that are produced from the AED system might contain isolated detections of events of very small duration or gaps of small duration between detection of events of the same type, which are usually erroneous. For instance, detection of a speech segment of duration less than 100 *msecs* is probably erroneous since there isn't a word, or even a phoneme with so sort duration. Moreover, several detections of an event that lie in the same neighbourhood, located very close to each other, should be grouped together to form a global event detection.

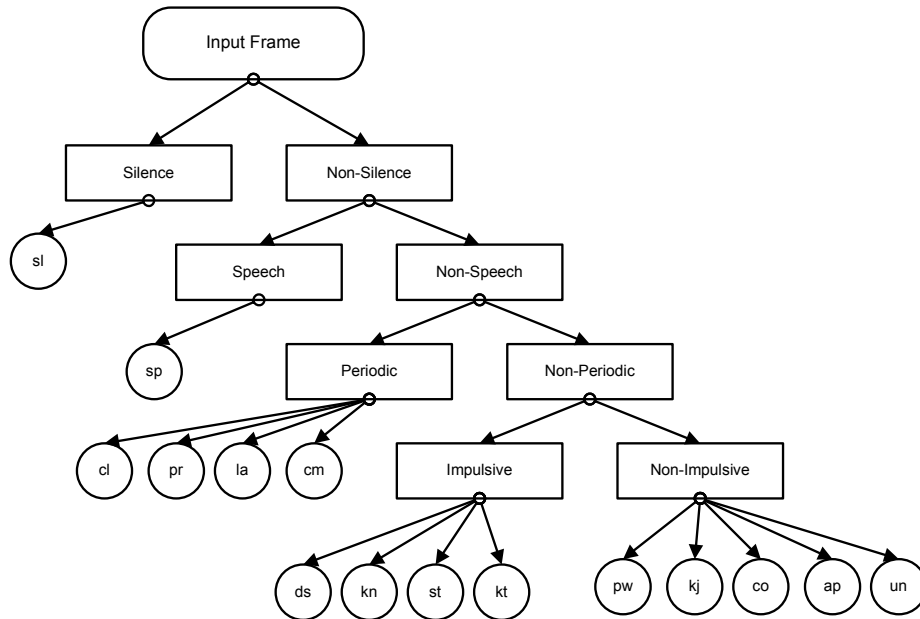


Fig. 4. A multiple-layer hierarchical structure for AED using HMMs.

For the smoothing of the raw detections a simple Markov model was employed, that was inspired from the hang-over schemes that are used for the smoothing of the detections of VAD systems. The basic idea of this post-processing system is to

- group together detections of the same event that are less than 3 frames far away from each other and
- to remove detections that after the grouping have length less than 5 frames.

It was found out that this smoothing procedure improves significantly the performance of the detector at since many erroneous mis-detections are removed, at the cost of dropping some accurate decisions though.

5 Results and Discussion

The performance of the system was evaluated by applying it to the testing data. These were spatially averaged, down-sampled, and separated into frames as described in section 2. From these frames feature vectors were extracted that were fed to the AED system. Finally the decision were smoothed with the use of a Markov model that imposed time duration constraints, as explained in section 4 and the smoothed decisions were converted into time-stamps. The results that were obtained are presented in Table 3.

Table 3. Evaluation results for AIT’s AED system

Primary System	
Accuracy	4.4%
Precision	3.4%
Recall	6.0%
Contrast System	
Accuracy	5.5%
Precision	4.4%
Recall	7.5%

The performance of this system is poor, contrary to what has been observed from the development data, where the recall metric was between 60% and 70%. Moreover, the performance of the contrast system is better than that of the primary, which was not the case when evaluating the performance of the system with the development data. The reasons for this are probably

- The fact that fusion of the data was performed through spatial averaging. Maybe it would have been more appropriate to derive a decision from each employed microphone and then fuse the decisions.
- The use of a small part of the isolated events databases data for training and from the testing data for the extraction of decisions (only the data obtained from microphones 13, 23, 33, 43, 53 of the mark III microphone array were used). This approach was applied in order to increase the coherence between testing and training data, but obviously it did not work.
- The use of a single Gaussian instead of Gaussian mixtures for the modelling of the continuous observation distribution of the states.
- The down-sampling which resulted in dropping information that could have been valuable for more accurate detections
- The choice of the feature vectors.

6 Conclusions

In this paper the Acoustic Event Detection (AED) system that was developed in Athens Information Technology (AIT) is presented. The components of this system, which are the pre-processing system, the detection system and the post-processing scheme are discussed into detail. The use of Hidden Markov Models arranged into a hierarchical structure is also presented in order to perform more robust detections. Finally the performance of the system is presented and various aspects that might possibly improve its accuracy are discussed.

7 Acknowledgements

This work is sponsored by the European Union under the integrated project CHIL, contract number 506909.

References

- [1] L.R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, vol .77, no. 2, pp. 257-286, 1989.
- [2] C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 2, 121–167, 1998.
- [3] 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, Lecture Notes in Computer Science, April 2006.
- [4] R.O. Duda P.E. Hart and D.G. Stork, Pattern Classification, John Willey & Sons, 2001
- [5] L. Mauuary and J. Monné, Speech/non-speech Detection for Voice Response Systems, in *Eurospeech'93*, Berlin, Germany, 1993, pp 1097–1100
- [6] A. Martin, D. Charlet and L. Mauuary, Robust Speech/Non-Speech Detection Using LDA Applied to MFCC, ICASSP, 2001
- [7] J. Ramirez, J.C. Segura , C. Benitez, L. Garcia, A. Rubio, Statistical Voice Activity Detection Using a Multiple Observation Likelihood ratio Test, IEEE Signal Processing Letters, vol. 12, no. 10, pp. 689–692, 2005.