An HMM Based System for Acoustic Event Detection

Christian Zieger*

FBK-irst,
Via Sommarive 18,
38050 Povo, Trento,
Italy
zieger@itc.it

Abstract. This paper deals with the CLEAR 2007 evaluation on the detection of acoustic events which happen during seminars. The proposed system first converts an audio sequence in a stream of MFCC features, then a detecting/classifying block identifies an acoustic event with time stamps and assign to it a label among all possible event labels. Identification and classification are based on Hidden Markov Models (HMM). The results, measured in terms of two metrics (accuracy and error rate) are obtained applying the implemented system on the interactive seminars collected under the CHIL project. Final not very good results highlight the task complexity.

1 Introduction

Acoustic scene analysis consists in describing all possible acoustic events in terms of space, time or type by means of a single microphone or a distributed microphone network that constantly monitors the environment [1]. In the CHIL project acoustic scene analysis has been adopted to describe automatically human interactions and interactions between humans and environment. This work focuses on the problem of detecting acoustic events, that is identifying an acoustic event with its timestamps and classifying it selecting among a list of predefined possible events.

In literature acoustic event detection has been studied in different fields: in [2] the authors focus on detecting a single event, in [3,4] speech and music classification is explored, in [5] acoustic events for medical telesurvey are considered, in [6] audio events are detected to automatically extract highlights from baseball, golf and soccer matches and [7] detection of animal sounds is investigated.

This paper addresses the CLEAR 2007 evaluation on acoustic event detection for seminars. In particular, the evaluation considers a list of 12 events (named CHIL events): door or table knock (kn); door slam (ds), steps (st), chair moving (cm), spoon clings or cup jingle (cl), paper wrapping (pw), key jingle (kj),

* This work was partially funded by the European Community under the CHIL and DICIT projects.

R. Stiefelhagen et al. (Eds.): CLEAR 2007 and RT 2007, LNCS 4625, pp. 338–344, 2008.
© Springer-Verlag Berlin Heidelberg 2008

keyboard typing (kt), phone ring/music (pr), applause (ap), cough (co), laugh (la). All recorded interactive seminars are acquired through a distributed microphone network composed by T-shaped arrays, the linear NIST markIII array and tabletop microphones. The proposed acoustic event detector (AED) considers the audio stream of a fixed single microphone, converts it in a feature vector sequence, on which an Hidden Markov Model (HMM) based detection stage is applied. The paper is organized as follows: section 2 will describe the inplemented AED system, section 3 reports on the evaluation results specifying the metrics used to evaluate system performance and finally the last section outlines some conclusions.

2 AED Based on HMM

The block diagram of the implemented AED consists in two blocks. The first one, the front-end, converts an audio stream into a sequence of acoustic parameter vectors. The second one, the event detector/classifier, identifies the acoustic events exploiting previously trained acoustic models.

2.1 Front-End

The audio signal sampled at 44.1 kHz coming from a fixed microphone is converted by the front-end in a feature vector stream of Mel Frequency Cepstral Coefficients (MFCC), which are widely used in the speech recognition field [8,9]. The Mel frequency equispaced triangular filter used are 24, while the cepstral coefficients extracted after the DCT operation are 12. The analysis window is 20 ms and the step of the sliding window is 10 ms. The hamming window is used. The first and second derivatives of the MFCCs are also computed and appended to the feature vector. The signal energy is not considered since the seminars are characterized by energy conditions that can change considerably from seminar to seminar and even during the seminar. Its first and second derivatives are appended to the acoustic feature vector instead. In conclusion, the acoustic feature vector is composed by 38 elements:

- 12 MFCCs + first and second derivatives
- first and second derivatives of the signal energy

In table 1 the system front-end parameters are summarized and in figure 1 the block diagram of the front-end is reported.

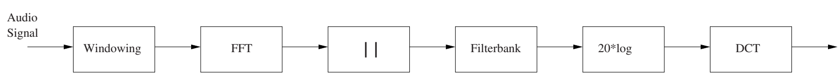

Fig. 1. Block diagram of the AED front-end

Table 1. Front-end parameters

front-end parameters	
sampling frequency	44.1 kHz
analysis window	20 ms
analysis step	10 ms
window type	Hamming
number of MEL filter bank channels	24
number of cepstral coefficients	12

2.2 Acoustic Event Detector

The acoustic event detector implemented is based on HMMs [8,9]. For every event in the list of all possible CHIL events an HMM is trained using the databases available in the development set, which is composed by the isolated acoustic event databases, collected by ITC and UPC (see [10] for a description), and by 4 interactive seminars whose characteristics in terms of length and number of events are reported in tables 2 and 3. The same event type can differ from the acoustic point of view from site to site according to the construction material of the object that produces the acoustic event (steps, chair moving, door slam). For this reason the not uniform distribution of the events among all seminars can give rise to a possible mismatch between acoustic models and test set. Further models for speech and silence are added to the list of all possible events in order to identify and consequently reject speech and silence sequences. The speech model is trained selecting from the interactive seminars those speech sequences that do not temporally overlap with other events.

The HMM topology adopted is left-right, with 3 states. The number of the Gaussian mixtures for each state is 128. Diagonal covariance matrix is adopted. The optimum events sequence is obtained applying the viterbi algorithm to the whole converted audio segment using the models previously trained. Site dependent (*SD*) and site independent (*SI*) systems are implemented. In the former a set of acoustic models is trained for each site using the corresponding training data. Then, in the detection step, the system selects acoustic models according to the prior information about the room type (UPC, ITC, UKA, AIT). Figure 2 reports on the block diagram of the *SD* AED. In the latter the same acoustic models, trained with all available databases, are used for each room (see figure 3) instead.

3 Evaluation Results

In this section the results of the evaluation for the *SD* and *SI* systems are reported. Let us notice that *SI* system results have been submitted after the official evaluation dead line, that is after having the ground truth. According to the evaluation plan, system performance is measured by means of two metrics, called AED-ACC and AED-ER.

Table 2. Number of occurrences of each event in single seminar

event type	ITC	UKA	AIT	UPC
kj	4	4	8	5
sp	304	435	195	190
pw	13	61	27	20
cl	7	0	0	22
kn	9	18	15	29
st	5	24	16	38
cm	48	123	16	43
kt	6	6	15	14
la	4	20	6	12
un	59	154	47	36
pr	6	0	8	4
co	11	20	4	16
ap	2	2	0	3
ds	23	4	18	11

Table 3. Seminar lengths in minutes

Seminar type	**length in minutes**
ITC	30
UKA	44
AIT	32
UPC	23

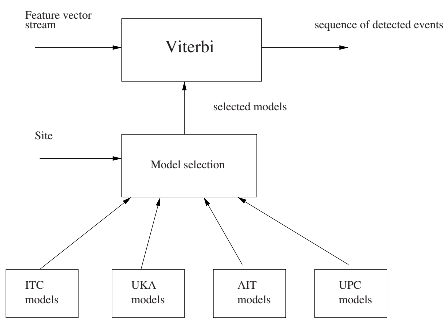

Fig. 2. Block diagram of the *SD* AED

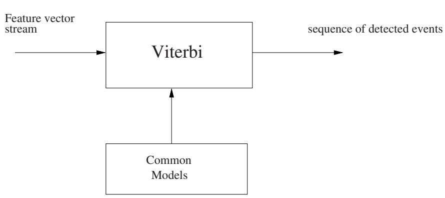

Fig. 3. Block diagram of the *SI* AED

3.1 Metrics

AED-ACC is defined as

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

$$Precision = \frac{\text{number of correct system output AEs}}{\text{number of all system output AEs}} \quad (2)$$

$$Recall = \frac{\text{number of correctly detected reference AEs}}{\text{number of all reference AEs}} \quad (3)$$

A system output acoustic event (AE) is considered correct if there exist at least one reference AE whose temporal center is situated between the timestamps of the system output AE and the label of the system output and the reference AE is the same, or if the temporal center of the system output AE lies between the timestamps of at least one reference AE and the label of the system output AE and the reference AE is the same. A reference is considered correctly detected if there exist at least one system output AE whose temporal center is situated between the timestamps of the reference AE and the label of the system output AE and the reference AE is the same, or if the temporal center of the reference AE lies between the timestamps of at least one system output AE and the label of the system output AE and the reference AE is the same. In this evaluation β is set to one.

AED-ER is a metric for measuring the temporal resolution of the detected event and is defined as follows:

$$AED{-}ER = \frac{\sum\limits_{\text{all seg}} \{dur(seg) * (\max(N_{REF}(seg), N_{SYS}(seg)) - N_{correct}(seg))\}}{\sum\limits_{\text{all seg}} \{dur(seg) * N_{REF}(seg)\}} \quad (4)$$

where $dur(seg)$ is the duration of the segment seg, $N_{REF}(seg)$ is the number of reference events in seg, $N_{SYS}(seg)$ is the number of system output events in seg, $N_{correct}(seg)$ is the number of reference events in seg which have a corresponding mapped system output AEs in seg.

3.2 Results

The evaluation test set is composed by 20 audio segments of 5 minute length. In table 4 the number of occurences, the total duration in seconds and the percentage of event shorter then 1 second are reported for each transcribed event in order to give an idea about the task complexity. First let us notice that the database is not balanced, in fact there are more frequent events like steps, chair moving, laugh, knock and events that happen rather seldom like applause, phone ring, cup jingle. This unbalancing is not considered in the metric for the system performance. Morover events not to be detected, like speech and unknown, are the most frequent events during the seminar, in fact they doubled the number

Table 4. Statistics in terms of number, duration, percentage of short events for each event in the evaluation test set

event	event number	total duration in s.	% of events shorter than 1 s.
kj	32	32.40	46.875 %
sp	1239	1241.60	19.53 %
pw	88	88.71	31.81 %
cl	28	29.35	39.28 %
kn	153	153.55	88.88 %
st	498	503.95	53.21 %
cm	226	232.64	57.07 %
kt	105	107.91	52.38 %
la	154	154.3	36.36 %
un	559	559.27	75.49 %
pr	25	26.08	40 %
co	36	36.77	75 %
ap	13	17.45	23.07 %
ds	76	76.51	46.05 %
total chil events	1434	1459.6	53.69 %
total non chil events	3232	1798.3	36.92 %

Table 5. Evaluation results in terms of AED-ACC, Precision, Recall and AED-ER for the site dependent system (results submitted before the evaluation dead line)

System type	AED-ACC	Precision	Recall	AED-ER
SD	23.4 %	35.4 %	17.5 %	109.07 %

of CHIL events and they temporally overlap with CHIL events. Observing the total duration of each event it can be noted that the average duration of every single event is about 1 second. Short events which are the most difficult to detect, represent about the 53 % of the total number of CHIL events. Short events that happen very often are knock, cough, chair moving, keyboard typing.

In tables 5 and 6 the results in terms of AED-ACC, Precision, Recall and AED-ER for both *SD* and *SI* systems are shown. The results confirm the difficulty of the task as previously mentioned. *SD* system guarantees better results in terms of AED-ACC, but the worst ones in terms of AED-ER. It can be useful to compare the results for each single site as reported in table 7. *SD* system reduces the mismatch between training and test data yielding in general better results for UKA and AIT, but worst results for UPC and ITC.

Table 6. Evaluation results in terms of AED-ACC, Precision, Recall and AED-ER for the site independent system (results submitted after the evaluation dead line)

System type	AED-ACC	Precision	Recall	AED-ER
SI	26.3 %	39.1 %	19.9 %	111.33 %

Table 7. Results for diffrent rooms and systems

Site	AED-ACC (SD)	AED-ACC (SI)	AED-ER (SD)	AED-ER (SI)
AIT	16.8 %	9.2 %	103.44 %	98.80 %
UKA	11.8 %	6.6 %	157.07 %	141.32 %
UPC	29.0 %	39.4 %	103.33%	116.25 %
ITC	30.0 %	32.0 %	86.93 %	93.97 %

4 Conclusions

In this paper an HMM based acoustic event detector with site dependent and site independent models has been introduced for the CLEAR 2007 evaluation on the acoustic event detection task. System performance have been measured in terms of two metrics using as test set 20 seminars, each 5 minutes long, collected in four rooms under the CHIL project. A description of the test database, characterized by very short events temporally overlapping with other disturbing events, like speech and unknown, let suppose the high difficulty of the considered task. The evaluation results have confirmed this hypothesis. Moreover the low results do not allow to make a good comparison between the *SD* and *SI*.

References

1. Wang, D., Brown, G.: Computational Auditory Scene Analysis: Principles, Algorithms and Applications. Wiley-IEEE Press (2006)
2. Kennedy, L., Ellis, D.: Laughter detection in meetings. In: NIST ICASSP Meeting Recognition Workshop, Montreal, Canada, pp. 118–121 (2004)
3. Lu, L., Hong-Jiang, Z.J.H.: Content analysis for audio classification and segmentation. IEEE Transaction on Speech and Audio processing 10(7), 504–516 (2002)
4. Pinquier, J., Rouas, J.L., Andrè-Obrecht, R.: Robust speech / music classification in audio documents. In: Proc. ICSLP, Denver, USA, vol. 3 (2002) 2005–2008
5. Vacher, M., Istrate, D., Serigna, J.F.: Sound detection and classification trough transient models using wavelet coefficient trees. In: EUSIPCO, Vienna, Austria, pp. 1171–1174 (2004)
6. Xiong, Z., Radhakrishnan, R., Divakaran, A., Huang, T.: Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In: ICME 2003, Baltimora, USA, vol. 3, pp. 401–404 (2003)
7. Slaney, M.: Mixtures of probability experts for audio retrieval and indexing. In: ICME 2002, Ischia, Italy, vol. 1, pp. 345–348 (2002)
8. Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs (1993)
9. Rabiner, R.L.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
10. Temko, A., Malkin, R., Zieger, C., Macho, D., Nadeu, C., Omologo, M.: Clear evaluation of acoustic event detection and classification systems. In: Stiefelhagen, R., Garofolo, J.S. (eds.) CLEAR 2006. LNCS, vol. 4122, Springer, Heidelberg (2007)