Detection of Acoustic Events in Interactive Seminar Data with Temporal Overlaps

Andrey Temko and Climent Nadeu
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
{temko, climent}@talp.upc.edu

Abstract

In Acoustic Event Detection (AED), both the identity of sounds and their position in time have to be obtained. In this paper, we first present our SVM-based system for AED in real-time conditions, along with the databases and metrics developed for the international evaluation campaign CLEAR 2007. In that evaluation, which was carried out with a real environment database that consists of interactive seminar recordings, the biggest encountered problem for AED was the presence of temporal overlaps, since they account for more than 70% of errors. In this paper we also report an initial attempt to deal with the overlap problem at the level of models. A two-step approach is proposed and it is tested firstly with artificially–overlapped acoustic data, and then with the above-mentioned seminar data.

Index Terms: acoustic event detection, temporal overlaps, support vector machines

1. Introduction

In our AED work we aim at processing the acoustic signals collected by distant microphones in meeting-room or classroom environments to convert them into symbolic descriptions corresponding to a listener's perception of the different sound events that are present in the signals and their sources. The detection of the acoustic events (AE) that are naturally produced in such environments 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.

Unlike, acoustic event classification [1], in AED the classification task is not performed on previously-segmented AEs so that both the identity of sounds and their position in time have to be obtained. Although the task of AED is relatively new, it has been already evaluated for meeting-room environments in the framework of two international evaluation campaigns organized by the project CHIL [2]: in CLEAR 2006, by 3 participants, and in CLEAR 2007, by 6 participants. The results (in CLEAR 2006) with databases of isolated AEs [3], where AEs do not have temporal overlaps, were rather good; however, a large decay in performance was observed (in both CLEAR 2006 and CLEAR 2007) when evaluating the developed systems with a real environment database that consists of interactive seminar recordings. The biggest problem for AED in that real environment appeared to be overlaps, i.e. temporal intervals where the AE of interest is overlapped with speech and/or other AEs. It was found that the overlapping segments account for more than 70% of errors produced by every submitted system [4]. Actually, the problem of acoustic overlaps is closely related to the "cocktail-party" problem [5]. In the latter, however, one usually tries to separate one speech source from others. Conversely, in AED we would like to separate acoustic events from speech. A related case is an overlap of different speakers which has been addressed by the NIST RT-07 [6] evaluation campaign, since the tasks (e.g. Speaker Diarization) have been evaluated with signal segments that include speaker overlaps.

Conceptually, the problem of overlaps can be addressed at different system levels. At the signal level, it can be dealt with source separation techniques like ICA [7]. At the level of decision, different weights can be assigned to particular microphones within the multi-microphone system architecture, assuming that the audio sources (in our case, speech and acoustic events) are well separated in space, and that a given acoustic event may correspond to the most powerful signal in some microphone.

In this paper, we first present our SVM-based system for AED, along with the databases and metrics developed for CLEAR 2007 evaluations. Then, an initial attempt to deal with the overlap problem at the level of models is reported. A two-step approach is proposed and tested firstly with artificially–overlapped acoustic data, and then with the above-mentioned seminar data.

The rest of the paper is organized as follows. Databases and metrics are presented in section 2. The various tested systems are described in Section 3. Finally, Section 4 includes the experimental results and a discussion.

2. Databases and metrics

The 12 evaluated 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. The database used in the CLEAR evaluation campaign 2007 consists of two databases of isolated acoustic events (UPC-iso and ITC-iso columns in Table 1) [3] and 25 interactive seminars of approximately 30 min long each that have been recorded by 5 CHIL partners 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 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 67% of time, the AEs are overlapped either with “Speech” or other AEs.

The aim of the detection accuracy (AED-ACC) metric is to score detection of all instances of what is considered 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):

$$AED - ACC = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (1)$$

where

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

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

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.

3. SVM-based AED systems

3.1. 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 [1]: 1) 16 frequency-filtered 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. In both training and testing processes, the mean and the standard deviation parameters have been computed over all frames in a 0.5-second window with a 100ms shift (which empirically were found to be the optimal window length and shift) as it is shown in Figure 1, thus forming one vector of 120 elements.

3.2. Classifiers

Three SVM-based systems have been developed in order to find a way to increase the system performance, especially in the time intervals where overlap is present. No post-processing (i.e. smoothing, hangover, etc) was applied to the outputs of the designed systems, and only one Mark III microphone channel is used throughout the work. Concerning SVM training, the standard operations are undertaken: anisotropic data normalization with the normalization templates that are applied afterwards to test data, and 5-fold cross-validation to obtain optimal values of both the Gaussian kernel parameter and the C parameter.

As baseline, we use a system similar to the one implemented and tested in the CLEAR evaluations in 2007 [4]. In that system (hereafter referred as baseline system) SVM classifiers have been trained using the isolated AEs from the two databases of isolated acoustic events, along with segments from the development data seminars, that include both isolated AEs and AEs overlapped with speech. The segments that contain the overlaps of two or more AEs, either with or without speech, are not used. The 1 vs. 1 DAG multi-class strategy [8] has been chosen to classify among 14 classes that include “Speech”, “Unknown”, and the 12 evaluated classes of AEs. Besides, a “Silence” vs. “Non-silence” SVM classifier has been trained, where “Non-silence” class includes all 14 classes.

Segments with overlapped sounds are included in the corresponding classes. To see the contribution of the overlapping segments to the baseline performance we have trained a version of the system only with isolated instances. That system (hereafter referred as ISO system) is supposed to show better performance than the baseline system on audio

![Figure 1. SVM-based AED system](image-url)
segments where the acoustic events appear isolated while its performance on whole seminars may be worse than the baseline due to a worse classification of overlapped segments, since they were not modeled during training.

In this work, in order to deal with the problem of overlaps, we propose a two-step approach. Firstly, instead of including the segments with overlaps in the corresponding classes, as it is done in the baseline test, we define another class (named [mp]) that contains all overlaps, and we use it with the ISO system to train the SVM classifiers which are needed to complete the set of 1-vs-1 SVM classifiers. Then, in order to classify the detected overlaps using the defined AE classes, we employ the confusion-based clustering scheme proposed in [1]. In that scheme, an optimal decision, tree with an SVM at each node, is computed from the confusion matrix obtained with development data. The confusion matrix is used to find the best way of splitting the set of classes at a given node into two clusters with the least mutual confusion. In [1], the confusion-based clustering has been exploited partially to introduce optimal feature sets on each step of classification. In this work we benefit from another characteristic of the method. Specifically, this approach (hereafter referred as ISO-CLUSTER) naturally enables to include, along with the isolated AEs, segments that contain two or more overlapped AEs, either with or without speech, and which were not used to train the baseline and ISO systems. For example, the cluster that consists of the classes [pw], [kt] and [st] from Figure 2 is modeled using the isolated AEs instances, AE data of the same classes that are overlapped with speech, and also other audio segments where those 3 AE classes are overlapped among themselves. The proposed ISO_CLUSTER scheme is presented in Figure 2.

4. Results and discussion

4.1. Experiments with artificially–overlapped data

The proposed systems of AED are firstly tested with acoustic events artificially overlapped with speech. For that purpose, a small database of isolated acoustic events was recorded in the UPC smart-room, along with speech segments, using a quiet and controlled environment (see Table 1, ISO-toy column). The artificially-overlapped acoustic database was constructed by adding speech segments to acoustic events using several signal-to-noise ratios (SNR) (speech is considered the "noise" in this experiment, and the SNR is measured only in the overlapped portion of the acoustic event), using several degrees of overlap between speech and acoustic events, and adding speech to either the beginning or the end of the time segments of the acoustic events. The constructed database is used solely for testing with the models trained on development data from Table 1. The results are shown in Figure 3.

As we can see, the ISO-CLUSTER system obtains higher accuracy than the other two systems for almost every SNR and overlap condition, being the differences more remarkable for lower SNRs and lower percentages of overlap. It can be also seen from Figure 3 that the lower the SNR and the higher the percentage of overlap, the higher the difference in
performance between both the baseline and the ISO-CLUSTER method, which model overlaps, and the ISO method, where overlaps are not used in training. However, as expected, for the same reason, for high SNR values, the ISO system performs consistently better than the baseline. The proposed ISO-CLUSTER can be thought as an approach that performs as well as the ISO system on data where overlaps do not have much influence (high SNR) and where no [mp] labels are produced, but performs better than the ISO system when the overlap effect starts to be more noticeable (low SNR and/or high percentage of overlap). In the worst case, with SNR=-10dBs and 100% of overlap, the performance of all systems is low, and the ISO system performs much worse than the other two systems.

For the sake of space we do not present the results for right-side overlap. However, from the results we saw that the still best-performing ISO-CLUSTER system shows in general a lower accuracy for right-side speech contamination than for the left-side one, especially for low degrees of overlap. This observation may indicate that the beginning parts of acoustic events are not as important as the ending parts. In fact, e.g. for sounds like [ds] and [co] the beginning parts which are “door opening” and “inspiration”, respectively, may not bring any additional discriminative information. However, the performance of the other systems could not consistently confirm this statement.

Calculating the average performance of each system for all SNR levels, percentages of overlap and positions, in terms of AED-ACC, we have 81% for the ISO-CLUSTER system, 76% for the baseline, and 74% for the ISO system. According to the results obtained in this experiment, 21% error reduction obtained by the ISO-CLUSTER system with respect to the baseline is due to its ability to perform in a similar way to the other two systems for every condition.

4.2. Experiments with seminars recordings

Figure 4 shows the results obtained by applying the developed systems to the testing seminar database presented in Section 2. The final results of the systems depend on how the “sp”-vs-AEs classifiers are trained, since the AE classes can be favored more or less than speech. For this reason, Figure 4 shows the ROC curves on the precision and recall axis. As it can be seen from Figure 4, the proposed ISO CLUSTER two-step technique performs better than the baseline and ISO techniques. On the other hand, by considering the [mp] class which contains all overlaps, and transforming the reference transcriptions in order to include it, it is possible to calculate the maximum accuracy achievable when dealing with overlaps at the model-level with the first-step ISO system, i.e. assuming all detected overlaps are classified correctly at the second step. E.g. if class [st] is overlapped with class [sp], in the transformed reference transcriptions the overlapping segment is assigned to the [mp] label. Then the output of the system is compared to the transformed transcription so the hypothesis [mp] labels are collated with the reference [mp] labels without knowing the underlying overlapped events. The matching of the [mp] labels is counted as a correctly detected AE. In terms of the reference transcription, this implies that the underlying AEs from the considered overlapping segment have been correctly classified.

The results obtained with that non-practical approach, hereafter referred as ISO ORACLE, are used as a reference. From Figure 4 we can see that, although the clustering scheme used for [mp] class re-classification leads to some improvement, the ISO ORACLE results are much better than those from the ISO and ISO-CLUSTER systems. This fact indicates that detecting the presence of overlapped sounds may largely help to increase the AED performance.

The big points in Figure 4 indicate for the combinations of precision and recall that produce the highest AED-ACC scores. It is worth mentioning, that the best accuracy of 38.5% obtained with ISO CLUSTER system surpasses the best score of the systems submitted to the CLEAR 2007 evaluations (that was 33.6%). Interesting enough, by comparing the best score on seminar data with the results on artificially-overlapped data we conclude that the seminar data corresponds to a situation which is worse than 100% overlap of isolated AEs recording (including speech), at -10 dB power ration between AEs and speech.

Future work will be oriented to deal with overlaps at the decision level with multiple microphones.

5. Acknowledgements

The authors want to thank Khalid Daoudi (IRIT-CNRS) for fruitful discussions and useful suggestions about the subject during the stay of the first author in Toulouse. This work has been partially funded by the Spanish Government project SAPIRE (TEC2007-65470).

6. References