ACOUSTIC DETECTION AND CLASSIFICATION USING TEMPORAL AND FREQUENCY MULTIPLE ENERGY DETECTOR FEATURES

J. Moragues, A. Serrano, L. Vergara and J. Gosálbez

Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM)
Polytechnic University of Valencia (UPV), Camino de Vera, s/n, 46022 Valencia, Spain
{jormoes, arsercar}@upvnet.upv.es, {lvergara, jorgocas}@dcom.upv.es

ABSTRACT

The problem of acoustic detection and recognition is of particular interest in surveillance applications, especially in noisy environments with sound sources of different nature. Therefore, we present a multiple energy detector (MED) structure which is used to extract a new set of features for classification, called frequency MED (FMED) and combined MED (CMED). The focus of this paper is to compare these two novel feature sets with the commonly used MFCC and to evaluate their performance in a general sound classification task with different acoustic sources and adverse noise conditions. The promising results obtained show that, in low SNR, the proposed CMED features work significantly better than the MFCC.

1. INTRODUCTION

There are a lot of areas in which the detection and classification of sound sources is required. Some of the most interesting ones are the surveillance applications in which the use of audio sensors is becoming increasingly important [1]. However, much of the recent work in this research area does not take into account the presence of the background noise. This normally decreases the performance of the detection and the classification phases, leading to an increase of the false alarms and to a poor recognition rate.

In real acoustic applications, as the sound sources are not completely known, the design of an appropriate detector is more difficult and energy detection is of interest. As we do not know the duration of the acoustic event, a multiple energy detector (MED) structure matched to different time durations can be used in order to fit the window size of the detector to the length of the novelty [2]. In the classification phase, most of the earlier studies present the Mel-Frequency Cepstral Coefficient (MFCC) features to be the most suitable for speech and sound sources identification [3]. Usually, MFCC offers a good performance but it is generally desired to have better features for noisy environments. In this case, the detector presented is able not only to determine the presence of an acoustic event inside a background noise, but also to provide information about it, which can be employed for classification. By using the signature that one event produces when it is processed by the MED, some appropriate novel features can be extracted in order to train a Bayesian classifier. In [4], the TMED features were presented and in this paper, they are used in combination with a new set of features extracted from a frequency MED structure.

This paper is organized as follows. Section 2 presents the principles of acoustic detection and the multiple energy detector structure used. Section 3 introduces the classification approach and the feature extraction process. In Section 4, the experimental setup is presented. Finally, the achieved results and the conclusion of our work are given in Sections 5 and 6.

2. DETECTION OF ACOUSTIC EVENTS

2.1. Multiple energy detector (MED)

One common method for detection of unknown signals is the energy detector (ED) which measures the energy in the received waveform over a specified observation time [5]. More formally, energy detectors are optimum solutions when the signal and the background noise are considered zero-mean multivariate Gaussian random vectors with uncorrelated components. The optimum test is:

$$\frac{\mathbf{y}^T\mathbf{y}}{\sigma_{\mathbf{w}}^2} \underset{\mathbf{H}_0}{\overset{\mathbf{H}_1}{\gtrless}} \lambda, \tag{1}$$

where $\mathbf{y}$ is the observation vector, $\sigma_{\mathbf{w}}^2$ is the noise variance and λ is the threshold that is set for a specific probability of false alarm (PFA).

However, there is an issue which must be considered for the practical application of EDs. As we ignore a-priori the novelty duration, we do not know the most appropriate size N of the observation vector $\mathbf{y}$ for implementing the detector. This question is addressed in [2] leading to a method based

This work has been supported by the Spanish Administration and the FEDER Programme of the EU under Grant TEC 2008-02975; and by the "Generalitat Valenciana" under Grant PROMETEO/2010/040.

on using multiple ED matched to multiple novelty durations. Taking into account this consideration, a multiple energy detector (MED) structure formed by several EDs with different sample size N of $\mathbf{y}$ is used to improve the acoustic time event detection of a single ED.

Since actual bandwidth of the event will generally be unknown too, the MED is also used in the frequency domain. Thereby, the presence of signal is decided if at least one of the ED decides it in the time or frequency structures. Other decision strategies could be devised, but since this is not he purpose of this paper, we will select the simplest one.

2.2. Proposed MED structure

Many different strategies for partitioning the initial observation vector could be used, but in the absence of any a-priori information we will consider L layers of partitions. Each detector u_{lm} corresponds to the output of the ED in m-th partition of level l, where $m = 1, \ldots, 2^l$ and $l = 0, \ldots, L-1$. At level 0 (top level) we have the original interval of N samples. In level 1, we have 2 non-overlapped intervals of $N/2$ samples each and so on until $L-1$ levels of successive divisions by 2. This produces a partition like the one represented in Fig. 1.

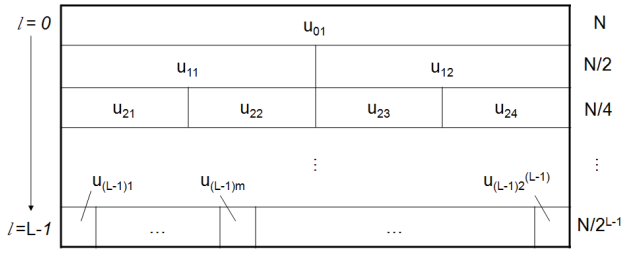

Fig. 1. MED structure with $L-1$ levels.

We assume that the same PFA is set for all levels, and this implies that a different threshold will be required for each N_l. In every selected interval, an ED of the form given in (1) is implemented considering vector $\mathbf{y}_l$, which is a portion of the initial observation vector $\mathbf{y}$. The threshold λ_l corresponding to the l-th ED can be obtained from:

$$\text{PFA}_l = Q\left(\frac{\lambda_l - N_l}{\sqrt{2N_l}}\right), \tag{2}$$

where Q stands for the error function. It must be noticed the statistical dependence among the different detectors. Therefore, the derivation of the global PFA corresponding to the detection structure is not easily obtained as shown in [2].

3. CLASSIFICATION

The sound source detected by the temporal and frequency MED structure described in Section 2.1 is then classified in two stages as follows.

3.1. Pre-classification

Firstly, the sound source is pre-classified as impulsive (IM) or non-impulsive (NIM) event. This is done by measuring the length of the event counting the number of detections in the lowest level of the temporal MED. An event is handled as impulsive if the total time duration of all detections in the time interval amounts less than a given duration.

We must point out the importance of the pre-classifier since further classification steps can take advantage of this decision. Thereby, when a new event is detected, only the categories belonging to the pre-classified class will be considered and a different parameter setup will be adjusted.

3.2. Feature extraction

Previous to the recognition algorithm, an analysis of the signal to be classified is required to extract the feature vector. Two types of features are evaluated in order to compare the performance of the classifier: MFCC and MED.

3.2.1. Modified MFCC extraction

Mel-Frequency Cepstral Coefficients (MFCC) are derived from a cepstral transformation and they represent the power spectrum of an acoustic segment. However, when the classification is subsequent to the detection phase, it is of particular importance to exclude segments with no event information since the recognition rate can considerably decrease due to this fact. To face this problem, a robust approach is used to extract features only from active segments that contain the acoustic event. Thereby, taking advantage of the temporal MED structure, only the active detection segments in the lowest level will be considered to extract the MFCC features for the classification phase, although other levels of the structure could be also used.

3.2.2. MED features

The information provided by the MED structures can be used not only for detecting new sound sources but also for classifying them. Therefore, novel features are extracted from the time as well as the frequency MED structures. They are based on the energy and have the advantage of being noise-independent since the MED continuously adapts to the noise changes over the time. They are calculated in the following way:

$$H(l) = \sum_{m=1}^{2^{l-1}} u(l,m) \qquad V(m) = \sum_{l=1}^{L} u\left(l, \left\lceil \tfrac{m}{2^{L-l}} \right\rceil\right), \tag{3}$$

where $H(l), \forall l = 1, \ldots, L-1$; and $V(m), \forall m = 1, \ldots, 2^l$, are the horizontal and vertical energy distribution of the event in the detection structure respectively.

Afterwards, the coefficients $H(l)$ and $V(m)$ are concatenated in a single feature vector for each MED structure and Principal Component Analysis (PCA) is applied in order to reduce the dimensionality. The features extracted from the time domain are called TMED, and the ones extracted from the frequency domain are denoted as FMED. In the same way, we introduce the combination of this two features sets leading to a new one referenced as CMED. In this case, the feature extraction method used is much less time consuming than for the MFCC features since only sums in the two dimensions of the MED are required.

3.3. Acoustic event modeling

The acoustic events are distinguished on the basis of their specific feature vectors using a Bayesian classifier [6]. Using a supervised approach for the training phase, a parametric model is calculated for each class. Then, a Bayes decision rule is applied in order to assign one of the pre-trained classes to each new sound source.

Assuming K classes, noted $c_i, i = 1, \ldots, K$, the posterior probability that a feature vector $\mathbf{x}$ belongs to a certain class c_i can be calculated using the Bayes theorem as:

$$P(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i)P(c_i)}{\sum_{j=1}^{K} p(\mathbf{x}|c_j)P(c_j)}, \quad (4)$$

where $P(c_i)$ is the class prior of c_i. Then, the acoustic event is assigned to the class c_i that satisfies $i = arg\left\{\max_i P(c_i|\mathbf{x})\right\}$. In this work, $P(c_i)$ are assumed to be equal since there is not any knowledge about the occurrence of the events.

The feature vectors have been modeled using a multivariate Gaussian probability density function, where the sample mean and the sample covariance matrix are the only parameters to be calculated for every class during the training process. Also Gaussian Mixture Model (GMM) has been tested obtaining worse classification results. This is related to the fact that the data test are very different due to the variety of SNR simulated and the GMM overfits the training data.

4. EXPERIMENTAL SETUP

Various acoustic events of different nature and duration, that can be indicative of dangerous situations in surveillance application, were recorded. Impulsive sound sources like hammer blows, knocks, breaking glasses and simulated shots were generated. Additionally, human speech, siren and metallic sounds were also analyzed as non-impulsive events. Furthermore, different SNR were performed by adding correlated Gaussian noise.

Real recordings were carried out in a typical office room using a multichannel audio data acquisition unit with a sampling frequency of 24 kHz. The microphones distribution used in the recordings consisted of two arrays of four omnidirectional microphones each and separated 1.9 meters. Each array of microphones is roughly an inverse t-shape geometry with a total width of 30 cm. Approximately 3 minutes of data were acquired for each sound source and each room position leading to a total amount of 600 acoustic events for each SNR. A MED structure of 9 levels was implemented, leading to a total duration of 5.46 seconds in the highest level and to an energy detector of 256 samples (≈10 ms) in the lowest one. The PFA was set to 10^{-8}.

To calculate the MFCC features, the signal is divided into frames of 21ms and they are characterized by a 13-dimensional vector. The first element of the vector corresponds to the DC component and it is not used for the classification. The number of features used for TMED and FMED depends on the pre-classification result. After applying PCA, a 14-dimensional and a 40-dimensional vector for the impulsive and the non-impulsive events are used respectively.

5. EVALUATION RESULTS

Several experiments have been carried out in order to test the performance of the detection and classification of acoustic events in presence of simulated Gaussian noise. We used the database described in the previous section and the results of each phase is presented below for several SNR conditions.

5.1. Event detection

Table 1 shows the probability of detection (PD) for impulsive and non-impulsive sounds. The detections obtained for the SNR equal to 20dB are used as ground-truth and, as expected, it can be observed how the PD decreases with SNR. Furthermore, there are some impulsive events, like shot and glass, which are more robust to the noise and present a high PD even in low SNR. However, for the non-impulsive events, the PD decreases considerably since the beginning and ending of the events are masked by the noise and they are not detected.

5.2. Pre-classifier

The results of the pre-classifier are presented in Table 2, where the percentage of the correct classification for impulsive and non-impulsive events is shown. It must be noticed how it works better for IM events since the tails of the NIM sounds can be misclassified. This fact can be appreciated specially in low SNR where there are fewer detections and therefore the pre-classifier is more likely to decide IM.

5.3. Feature evaluation

In this section, the classification results are presented assuming that the detection and the pre-classification stages are correct. A comparison between four audio features sets (MFCC, TMED, and the novel FMED and CMED)

source	SNR(dB)						
	10	0	-2	-4	-6	-8	-10
knock	99.3	99.3	99.3	98.1	90.1	82.1	72.9
hammer	98.1	94.6	93.3	92.4	90.9	84.6	74.6
glass	100	100	100	100	100	100	100
shot	99.4	99.4	99.4	99.4	99.4	99.4	99.4
metal	95.3	93.6	93.1	92.4	92.7	87.7	80.2
siren	90.3	83.8	78.1	73.3	64.9	52.1	36.9
speech	94.8	92.7	90.4	89.1	84.8	77.2	65.9

Table 1. PD(%) of impulsive and non-impulsive events in several SNR conditions.

source	SNR(dB)					
	20	10	0	-4	-8	-10
IM	97.6	97.6	97.5	97.5	99.1	99.0
NIM	89.9	90.4	89.8	87.2	84.1	80.7

Table 2. Pre-classification results (%) of impulsive and non-impulsive events in several SNR conditions.

are evaluated in their ability to differentiate acoustic events. Features extracted from detections in 20dB are used in the training step while the other SNRs are used for testing.

The recognition rates obtained for the non-impulsive events using FMED, CMED and MFCC are above 96% in all SNR. However, the TMED features present worse results due to the fact that they only consider the temporal information of the sound source and this makes more difficult to distinguish between events of the same time duration. Then, they are not appropriate to classify events of the same nature.

Fig. 2 shows the results achieved when classifying the impulsive sounds using the different feature sets. The results show that the performance of FMED is better than TMED, but they are complementary in the sense that we can combine them in CMED to get even a better classification rate. If we compare this new features against MFCC, we can see that MFCC behaves slightly better in SNRs higher than 5dB, but as the noise conditions become worse its performance considerably decreases. In this case, the CMED features reach better recognition rates and an improvement up to 15% can be obtained. Therefore, it can be seen that they are more robust to the noise conditions.

6. CONCLUSION AND FUTURE WORK

In this paper, the detection and classification of acoustic events in noisy environments are considered using a MED structure and a two-stage classification approach respectively. The results show the performance of the detection structure and the importance of the pre-classifier. Furthermore, two new set of features, denoted as FMED and CMED, were evaluated and compared with the TMED and MFCC. The experimental results show that MFCC performs slightly better in good SNR conditions. On the contrary, CMED presents a significant improvement of the classification accuracy of up to 15% in low SNR, especially for impulsive sounds. Further investigations will consider the possibility to combine both in order to improve the reliability of the classification.

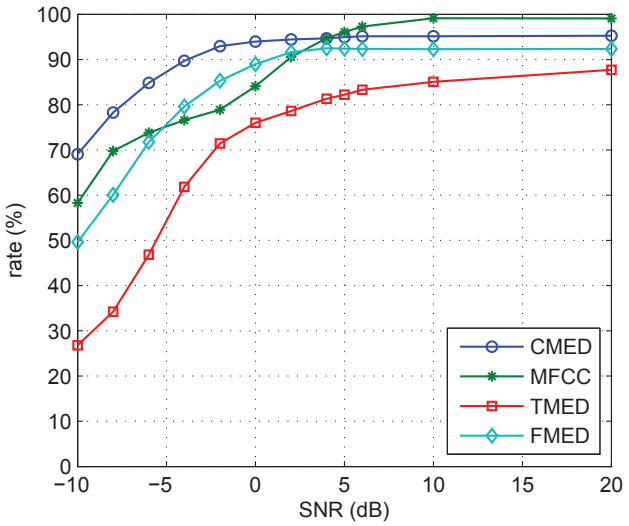

Fig. 2. Probability of classification for impulsive events, in several SNR conditions.

7. REFERENCES

[1] C. Zieger, A. Brutti, and P. Svaizer, "Acoustic based surveillance system for intrusion detection," in *Proc. of IEEE Int. Conf. on Advanced Video and Signal based Surveillance (AVSS'09)*, Genoa, Italy, 2009, pp. 314–319.

[2] L. Vergara, J. Moragues, J. Gosálbez, and A. Salazar, "Detection of signals of unknown duration by multiple energy detectors," *Signal Processing*, vol. 90, no. 2, pp. 719, 2010.

[3] A. Dufaux, *Detection and recognition of impulsive sounds signals*, Ph.D. dissertaion, Faculté des sciences de l'Université de Neachtel, Neuschatel, Switzerland, 2001.

[4] A. Swerdlow, J. Moragues, T. Machmer, L. Vergara, J. Gosálbez, and K. Kroschel, "Acoustic detection and classification of sound sources using temporal multiple energy detector features," in *Proc.* 17^{th} *European Signal Processing Conf. (EUSIPCO'09)*, Glasgow, Scotland, aug 2009.

[5] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, NJ: Prentice-Hall, 1st edition, 1998.

[6] R. O. Duda and P. E. Hart, *Pattern classification*, Wiley Interscience, 2nd edition, 2001.