Improving detection of acoustic signals by means of a time and frequency multiple energy detector

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

Abstract—**Standard energy detectors (ED) are optimum to detect unknown signals in presence of uncorrelated Gaussian noise. However, in real applications the signal duration and bandwidth are unpredictable and this fact can considerably degrade the detection performance if the appropriate observation vector length is not correctly selected. Therefore, a multiple energy detector (MED) structure is applied in the time as well as in the frequency domain and it is evaluated in real acoustic scenarios. The results obtained demonstrate the robustness of the MED structure and the improvements in performance reached in comparison to the standard ED.**

Index Terms—**multiple energy detector, acoustic event, frequency and time detection.**

I. INTRODUCTION

THERE are a lot of areas in which acoustic scene analysis is required. Some of the most interesting ones are the surveillance applications in which the signals recorded by a set of microphones are processed to extract as much information as possible of the environment [1], [2]. Other related works include acoustic event detection in order to determine the presence of sounds in real life scenarios [3]. Furthermore, there are some previous studies devoted to the evaluation of the detection performance in the frequency domain [4] as well as some applications (e.g, echo detection in radar, sonar or acoustics) where we may have an approximate idea of the signal assuming knowledge of its bandwidth [5]. However, much of the work in this research area does not take into account the actual duration of the signal and in consequence the segment length of the observation vector. In the context of novelty or event detection [6] any kind of signal is to be expected, hence signal duration and bandwidth incorporate an uncertainty to the problem since they are not available in practice. This fact normally decreases the performance of the detector, leading to miss detections and to an increase of the probability of false alarm (PFA).

The theoretical problem of detecting unknown signals with unknown duration has been previously presented in [7], where a multiple energy detector (MED) structure is implemented in order to match the different possible signal durations. No evaluation results of this novel approach have been performed in any real scenario or practical application, but the improvements provided by the theoretical results justify the utility and applicability to acoustic scenarios. Therefore, the focus of this work is to present a detailed evaluation in the framework of acoustic scene analysis where the multiple energy detector can provide a significant improvement in the detection of acoustic events. In particular, we study the performance of the MED not only in the time but also in the frequency domain since detection of signals with unknown bandwidth is also of particular interest. The Receiver Operating Characteristic (ROC) is used to summarize the robustness of the detectors tested in low signal to noise ratio (SNR) and in presence of real sound sources.

This paper is organized as follows. Section II presents the principles of acoustic detection using a MED structure for detecting signals of unknown duration and bandwidth. In Section III the experiments performed are presented and the different sound sources used are studied. Finally, the achieved results and the conclusion of our work are given in Sections IV and V.

The authors are with the "Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM)" in "Universidad Politécnica de Valencia", 46022 Valencia, SPAIN (e-mail: jormoes@upvnet.upv.es; Tel:+34963877308; Fax: +34963877919).

II. DETECTION OF ACOUSTIC SIGNALS WITH UNKNOWN DURATION AND BANDWIDTH

The detection problem is directly related to the knowledge of the acoustic signal that is to be detected. When sound sources are not completely known, the design of the appropriate detector is more difficult. In this case, one common method for detection is the energy detector (ED) which measures the energy in the received waveform over a specified observation time. Energy detectors are optimum solutions, for both Bayes and Neyman-Pearson criteria, for the following detection problem [8]:

$$\begin{array}{ll} H_0 : \mathbf{y} = \mathbf{w} & \mathbf{w} : N(0, \sigma_{\mathbf{w}}^2 \mathbf{I}) \\ H_1 : \mathbf{y} = \mathbf{s} + \mathbf{w} & \mathbf{s} : N(0, \sigma_{\mathbf{s}}^2 \mathbf{I}), \end{array} \quad (1)$$

where $\mathbf{s}$ is the unknown signal vector and $\mathbf{w}$ is the observed noise vector. In (1), both the noise and the signal are considered zero-mean multivariate Gaussian random vectors with uncorrelated components and variance $\sigma_{\mathbf{w}}^2$ and $\sigma_{\mathbf{s}}^2$ respectively.

The optimum test for (1) is:

$$\frac{\mathbf{y}^T \mathbf{y}}{\sigma_{\mathbf{w}}^2} \mathop{\gtrless}_{\mathbf{H}_0}^{\mathbf{H}_1} \lambda, \quad (2)$$

where λ is a threshold which depends on the required PFA. However, there is an issue which must be considered for the practical application of EDs. As we ignore a-priori the novelty duration and bandwidth, we do not know the most appropriate size N of the observation vector $\mathbf{y}$ for implementing the detector. This fact can be studied considering the probability of detection when N becomes large [8]:

$$PD \approx Q(Q^{-1}(PFA) - SNRN)), \quad (3)$$

where Q stands for the error function and $SNRN = SNR/\sqrt{2N}$ is a normalized signal to noise ratio with $SNR = \mathbf{s}^T\mathbf{s}/\sigma^2_{\mathbf{w}}$. Taking this into consideration and observing (3), we can conclude that for a given PFA, the PD not only depends on the SNR but also on the dimension N of the observation vector.

This question is addressed in [7] where the theoretical bases of a method consisting on using multiple EDs matched to different novelty durations is presented. Thereby, assuming that the observation vector $\mathbf{y}$ corresponds to N time samples, we consider L layers of partitions consisting on successive segmentation by a factor of 2 of the original observation interval (dimension N). In each layer l, we have 2^{l-1} EDs with a non-overlap observation vector of $N/2^{l-1}$ samples where $l = 1, \ldots, L$. Presence of signal is decided if at least one of them decides it. Therefore, we keep the simplicity of one ED but, on the contrary, we must consider the statistical dependence between the individual decisions at different layers. In this case, the theoretical performance of the MED can be studied deriving its probability of false alarm, denoted as PFA_{MED}, and the corresponding probability of detection. The PFA_{MED} can be obtained as:

$$PFA_{MED} = 1 - \frac{(1 - PFA)^{2^L - 1}}{(1 - Q(\sqrt{2} \cdot Q^{-1}(PFA)))^{2^{L-1} - 1}}, \quad (4)$$

where the PFA is the probability of false alarm of each individual detector and it is assumed to be the same in all layers ($PFA_l = PFA, \forall l$). However, for every layer it is required a different threshold λ_l obtained from:

$$\mathrm{PFA}_l = Q\left(\frac{\lambda_l - N_l}{\sqrt{2N_l}}\right), \quad (5)$$

where N_l is the observation vector size at layer l.

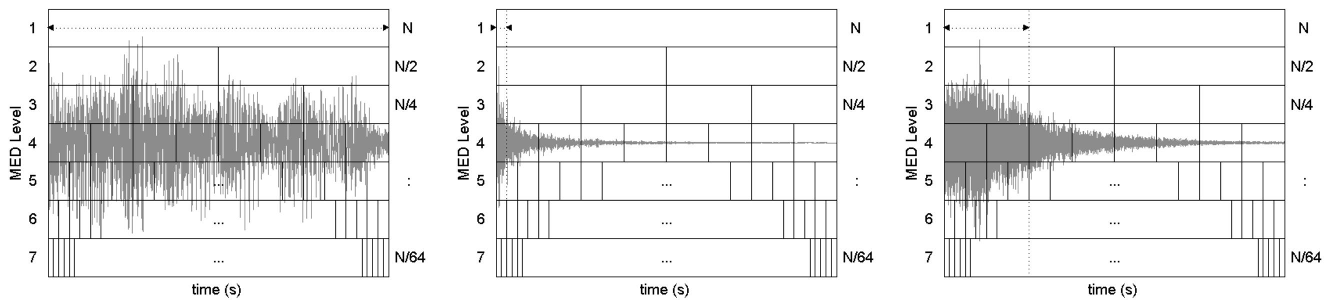
Furthermore, since the signals to be detected are completely unpredictable, the spectrum can also provide additional information. Thus, the detection problem in the frequency domain is of particular interest and it can be attempted by using the MED structure in order to detect signals of different bandwidth. In this later case, it is possible to use the same methodology previously described applying a fast Fourier transform (FFT) processor to the original observation vector $\mathbf{y}$ of dimension N (layer 1). Then, the same subdivision strategy as the one used in the time domain is applied. The detection problem is attempted at each individual frequency segment using the classical ED and computing the test statistic in a similar manner as in (2). In consequence, we consider both problems as conceptually equivalent and we will validate this technique in order to detect real signals.

III. Experimental Setup

The experiments have been performed with the objective of studying the performance of the MED in real acoustic scenarios. Various acoustic events of different nature and duration were recorded in a typical office room using a multichannel audio data acquisition unit with a sampling frequency of 24 kHz. The microphones distribution used consisted of two arrays, separated 1.9 meters, of four omnidirectional microphones with an inverse t-shape geometry and a total width of 30 cm. Approximately 3 minutes of data were acquired for each sound source at 3 different room positions, leading to a total amount of 600 acoustic events for each SNR. A time and frequency MED structure of 7 layers were used with an original observation vector in the highest layer of $N = 16384$ samples, leading to a total duration of 0.68 seconds and a total bandwidth of 12 kHz respectively. The PFA of the individual EDs was set to 10^{-8}.

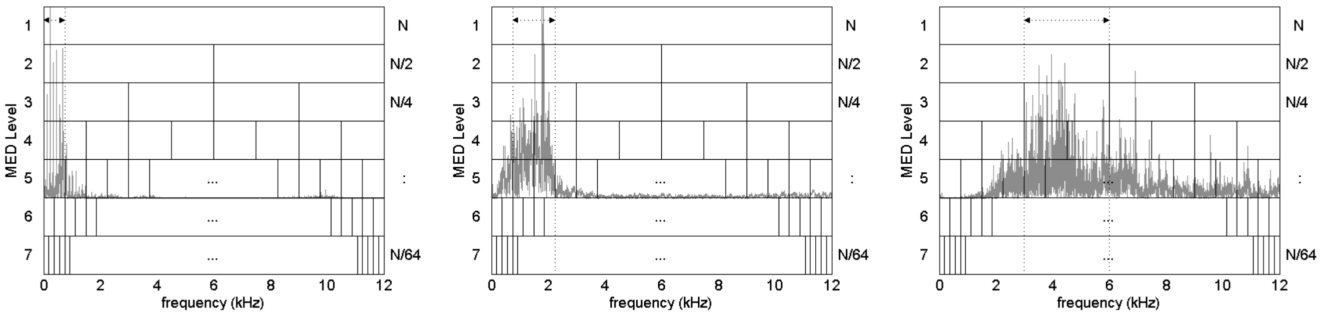
In order to contemplate the largest set of cases, we consider 3 acoustic events according to its time duration. Impulsive sound sources like *claps* and *breaking glasses* were generated, and additionally *human speech* was also analyzed as non-impulsive sound source. In Fig. 1, a time realization of the acoustic events is depicted superimposed on the time MED structure. As it can be observed, the *speech* signal extends its energy uniformly across the whole initial observation vector. On the contrary the *clap* signal concentrates its main energy across the first segment of layer 6 corresponding to $N/32$ samples. Finally, as an intermediate example between the foregoing two cases, the *breaking glass* signal extends its energy mainly through the first segment of the third layer.

In Fig. 2, we show the absolute value of the normalized frequency response of each acoustic signal, where again the frequency MED structure used is superimposed. As it can be observed, the acoustic signals selected have different spectrums and characteristics. The most important spectral infor-



(a) (b) (c)

Fig. 1: Realization of the acoustic signals superimposed on the time MED. (a) *Speech*, (b) *clap* and (c) *breaking glass*.



(a) (b) (c)

Fig. 2: Normalized frequency response of the acoustic signals superimposed on the frequency MED. (a) *Speech*, (b) *clap* and (c) *breaking glass*.

mation of the *speech* signal is mainly concentrated between 0 and 1 kHz. The *clap* example presents a low frequency spectrum mainly concentrated between 500 Hz and 2.2 kHz. The last one shows a totally different frequency response and corresponds to the *breaking glass* signal. In this case, the spectrum has higher frequency components and a wider bandwidth which is mainly extended between 3 and 6 kHz.

IV. Acoustic detection Results

In this section, the results of our experiments are presented and discussed. In Fig. 3 and Fig. 4, the time and frequency ROC curves of the MED are showed for different sound signals. In both cases, several $SNRN$ defined in the first layer of the structure ($SNRN = SNR/\sqrt{2N}$) were generated by adding white Gaussian noise, but only the most representative ones have been presented for each case. In addition, in the time as well as in the frequency domain, we compare the ROC curves considering several partitions of the original observation vector (dimension N), that is the number or layers used in the MED structure (L).

The main objective is to show and validate the interest of using the MED instead of the ED (only one layer, $L = 1$) when the signal duration or bandwidth are unknown. Thereby, we consider several cases in which the MED is used to detect signals with a time duration and bandwidth comparable to the segment length of different layers.

A. Signals with unknown time duration

In Fig. 3, we can observe the time MED performance when detecting 3 acoustic signals of different time duration. Fig. 3a shows the ROC curves for the *speech* signal which extends uniformly across the whole initial observation vector. In this case, the best performance is reached for $L = 1$ which is equivalent to the single ED. As expected, some degradation is obtained in the ROC when using unnecessary partitions ($L > 1$). On the contrary, as observed in Fig. 3b where a *clap* signal is to be detected, we obtain an improvement in the MED performance when increasing the number of layers used in comparison to the single ED ($L = 1$). The best results are obtained for $L = 7$ since the signal duration is more comparable to the observation segment of the bottom layer.

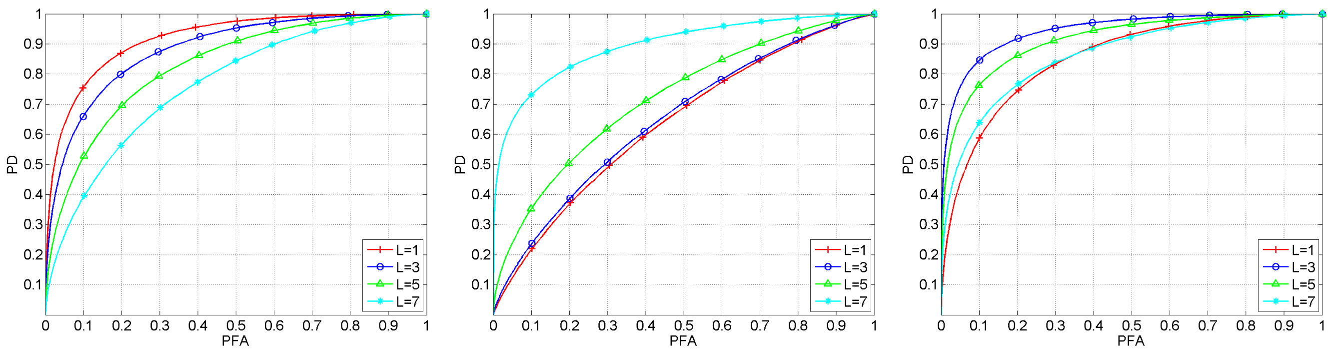
In Fig. 3c, we have an intermediate example where the acoustic signal has a time duration similar to the segment used in layer 3. As expected, the best ROC curve stands for $L = 3$. However, we must notice how the degradation suffered when using more layers ($L > 3$) is not as important as the degradation experimented when using no subdivisions ($L = 1$). Therefore, it is worthwhile to use as much layers as possible in the MED since the degradation suffered when we use unnecessary subdivisions is not as significant as the one experimented when they are not used.

B. Signals with unknown bandwidth

In Fig. 4, the experimental ROC curves obtained in detecting acoustic signals by using the frequency MED structure are presented. The first example is shown in Fig. 4a, where the best detection results are obtained for $L = 5$ since the *speech* bandwidth has a similar length to the observation segment of layer 5. The performance decreases when using more layers ($L > 5$), but as expected, the degradation is more significant when we use less layers ($L < 5$), specially for $L = 1$ which corresponds to the ED.

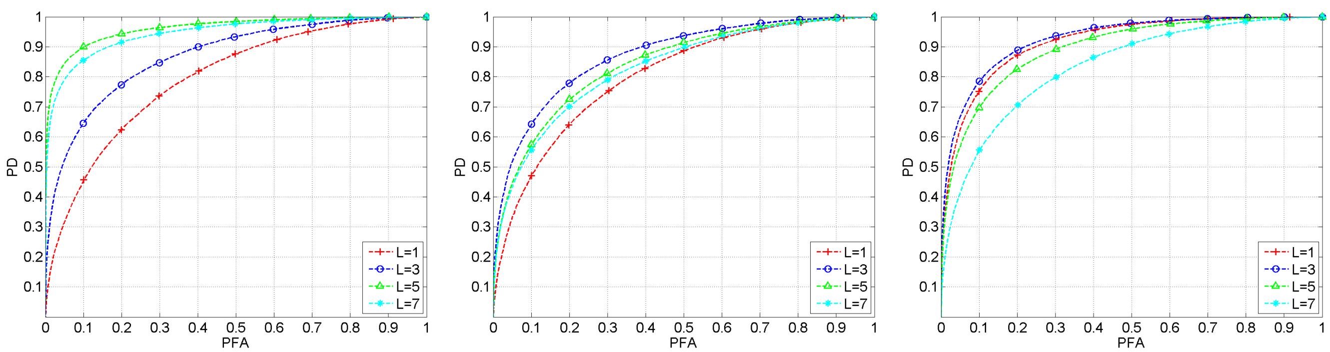
The other two examples are presented in Fig. 4b and Fig. 4c where a *clap* and a *breaking glass* signals are to be detected. In both cases, the optimum number of layers to be employed in the frequency MED structure is equal to 3 as observed. Comparing the spectrums of both signals (Fig. 2b and Fig. 2c), we observe how the *breaking glass* signal has a larger bandwidth and therefore a bigger number of optimum layers would be expected. However, this fact can be explained since in the *breaking glass* example the second best result is obtained for $L = 1$ while for the *clap* case is $L = 5$ according to its actual bandwidth.

It must be pointed out, that the frequency noise vectors are obtained taking the absolute value of the white Gaussian noise spectrum and therefore, they are Rayleigh distributed [8]. When the noise follows a non-Gaussian distribution, it is possible to implement other extensions of the ED, as presented in [9]. This would lead to a general improvement of PD for all ROC curves in Fig. 4, although the relative comparison among the different layers L would remain the same and similar results would be achieved.



(a) (b) (c)

Fig. 3: ROC curves of time MED structure using different layers (L). (a) *Speech* ($SNRN = 2$), (b) *clap* ($SNRN = 1$) and (c) *breaking glass* ($SNRN = 1.5$).



(a) (b) (c)

Fig. 4: ROC curves of frequency MED structure using different layers (L). (a) *Speech* ($SNRN = 1$), (b) *clap* ($SNRN = 1.2$) and (c) *breaking glass* ($SNRN = 2.5$).

V. Conclusion and Future Work

In this work, a novel detection structure denoted as MED is evaluated in real acoustic scenarios where the duration of the sound sources and their bandwidth are completely unknown. The MED is applied in the time as well as in the frequency domain and is based on multiple energy detectors with different time and frequency observation intervals respectively. A real data collection of acoustic events was carried out by recording signals of different nature and bandwidth. The experimental results were illustrated by means of the time and frequency ROC curves obtained for the MED structure in adverse noise conditions with low SNR. The study of the ROC curves showed the improvement in detection performance reached when using the MED structure ($L > 1$) in comparison to the single ED ($L = 1$) when the signal duration or bandwidth is smaller than the original observation vector of length N. In addition, since the signal duration and bandwidth are not known in advanced, it is showed how this improvement is worthwhile in spite of the possible degradation suffered when we use unnecessary layers of the MED structure. Further investigations will consider other possible partitions of the initial observation vector and various combined decisions of the time and frequency MED structures could be devised in order to improve the final detection process.

References

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proceedings of IEEE International Conference on Multimedia and Expo*, Amsterdam, Netherlands, 2005, pp. 1306–1309.

[2] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'09)*, 2009, pp. 165 –168.

[3] Antti Eronen Tuomas Virtanen Annamaria Mesaros, Toni Heittola, "Acoustic event detection in real life recordings," in *Proc.* 18^{th} *European Signal Processing Conf. (EUSIPCO'10)*, Aalborg, Denmark, aug 2010.

[4] Y.T. Chan, Q. Yuan, H.C. So, and R. Inkol, "Detection of stochastic signals in the frequency domain," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 37, no. 3, pp. 978–988, July 2001.

[5] E.D. Cheng, M. Piccardi, and T. Jan, "Stochastic boats generated acoustic target signal detection in time-frequency domain," in *Signal Processing and Information Technology, 2004. Proceedings of the Fourth IEEE International Symposium on*, 2004, pp. 429–432.

[6] M. Markou and S. Sameer, "Novelty detection: a review-part 1: statistical approaches," *Signal Processing*, vol. 83, pp. 2481–2497, November 2003.

[7] 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.

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

[9] J. Moragues, L. Vergara, J. Gosálbez, and I. Bosch, "An extended energy detector for non-gaussian and non-independent noise," *Signal Processing*, vol. 89, no. 4, pp. 656, 2009.