A Target Guided Subband Filter for Acoustic Event Detection in Noisy Environments Using Wavelet Packets

Zu-Ren Feng, *Member, IEEE*, Qing Zhou, Jun Zhang, Ping Jiang, and Xue-Wen Yang

***Abstract*—This paper deals with acoustic event detection (AED), such as screams, gunshots, and explosions, in noisy environments. The main aim is to improve the detection performance under adverse conditions with a very low signal-to-noise ratio (SNR). A novel filtering method combined with an energy detector is presented. The wavelet packet transform (WPT) is first used for time-frequency representation of the acoustic signals. The proposed filter in the wavelet packet domain then uses a priori knowledge of the target event and an estimate of noise features to selectively suppress the background noise. It is in fact a content-aware band-pass filter which can automatically pass the frequency bands that are more significant in the target than in the noise. Theoretical analysis shows that the proposed filtering method is capable of enhancing the target content while suppressing the background noise for signals with a low SNR. A condition to increase the probability of correct detection is also obtained. Experiments have been carried out on a large dataset of acoustic events that are contaminated by different types of environmental noise and white noise with varying SNRs. Results show that the proposed method is more robust and better adapted to noise than ordinary energy detectors, and it can work even with an SNR as low as −15 dB. A practical system for real time processing and multi-target detection is also proposed in this work.**

***Index Terms*—Acoustic event detection (AED), background noise, filter, wavelet packets.**

I. Introduction

ACOUSTIC event detection (AED) is a challenging and important research area for various applications, like security surveillance in public places [1], smart medical care for patients and the elderly [2], smart meetings [3], or robots working in adverse environments. In these detection systems, it is essential to keep the detection accuracy as high as possible, as any missed events may be costly. However, AED in real world environments is often faced with complex and noisy backgrounds. For one thing, a target acoustic event is likely to be drowned out by strong noise with a low signal-to-noise ratio (SNR). For another, highly non-stationary background noise could contain a large variety of acoustic events, and may merge with new events over time. Other similar but non-target sounds in the background may cause an increase of false detections. In this paper, we focus on robust AED under low SNR conditions.

The AED problem addressed in this paper is to identify the event of interest in audio signals and to determine its occurrence time [3]. A number of researchers have explored methods for AED in recent years. A simple but effective way for detecting unknown sounds is through a threshold based energy detector [4]–[7], often followed by a recognition step. Dufaux *et al.* [5] proposed to analyze energy variations of input audio signals, which were estimated in a fixed-size time window. Pulses of signals were detected by measuring the energy difference between the input and the output from a median filter. An errorless detection above 0 dB was achieved under white noise, but the performance deteriorated significantly when tested with real-world environmental noise. Istrate *et al.* [2] made an improvement by applying a pre-filtering technique where wavelet coefficients at upper levels were considered to be significant and summed as the energy. Better performance can be achieved for applications where the acoustic events to be detected are composed of high frequencies while the noise (e.g., in an apartment) has mainly low frequencies. Most recently, Ahmed *et al.* [6] proposed an analytical framework built on top of the detector proposed in [5], that allowed tuning of the threshold according to a given missed detection rate (MDR). It can optimize the trade-off between the missed detections at the detection stage and the computational load at the following recognition stage. In general, standard energy detector-based methods perform poorly under low SNR conditions and are often used as pre-processors to detect an abnormality from ambient noise.

Most other relevant research regards detection as a classification task using traditional machine learning techniques. The detection consists of two stages: frame-based feature extraction and classification [8][9]. Systems with different frameworks have been studied: hierarchical or nonhierarchical, supervised or unsupervised, online or offline. A typical gunshot detection system was proposed by Clavel *et al.* [10]. The input audio

Manuscript received July 14, 2014; revised October 21, 2014; accepted November 27, 2014. Date of publication December 18, 2014; date of current version January 15, 2015. This work was supported by the National Natural Science Foundation of China under Grants 61105034 and 61203350. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bozena Kostek.

Z.-R. Feng, Q. Zhou, J. Zhang, and X.-W. Yang are with the State Key Laboratory for Manufacturing Systems Engineering, Systems Engineering Institute, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: fzr9910@mail.xjtu.edu.cn; belief2012@stu.xjtu.edu.cn; zhangjun.jarry@stu.xjtu.edu.cn; michelyang1990@gmail.com).

P. Jiang is with the Department of Computer Science, University of Hull, Hull HU6 7RX, U.K. (e-mail: p.jiang@hull.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2381871

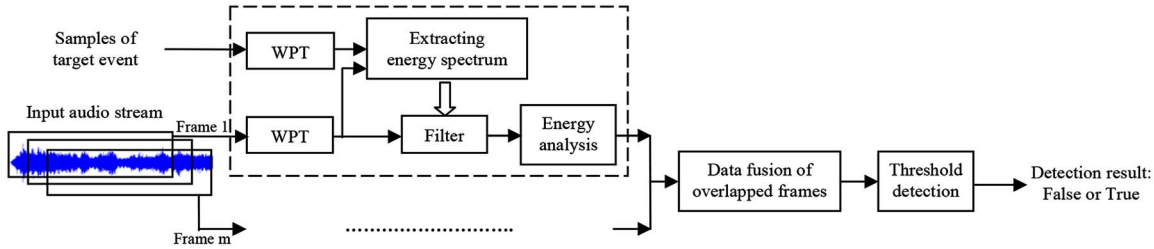


Fig. 1. Flowchart of the proposed filtering algorithm for AED.

stream was first segmented into short frames (20 ms) and various features in the time/frequency domain were captured. Then two Gaussian Mixture Models (GMMs) were trained with samples of gunshot and noise classes, respectively, resulting in a binary classifier to label each segment as a gunshot or normal. Acceptable results were reached, with a MDR of less than 11% and a false detection rate (FDR) less than 15% when the SNR was 5 dB or larger. A similar scheme was applied by Gerosa *et al.* [11] for gunshot and scream detection using two parallel GMM classifiers, achieving an accuracy of 90% and a false rejection rate of 8% under 0 dB. One obvious drawback of these supervised methods is that they are designed for specific applications and are highly reliant on a priori knowledge of the background noise, and thus cannot adapt to environments with strong dynamics. Ntalampiras *et al.* [1] presented a retraining methodology via a feedback loop, enabling the adaptation of the detection system to the surrounding environment. A hierarchical system based on Hidden Markov Models (HMMs) was also designed for detecting abnormal events (screams, gunshots, or explosions) in a subway environment. The proposed system demonstrated a good performance at the first stage for normal/abnormal classification, having an equal error rate (EER) of 8.53% at -5 dB, but it rises to 22.67% at the second stage due to confusion between the three categories.

In this work, a novel filtering method for AED is developed, with the aim of improving the performance, especially under adverse conditions with high-level background noise. The proposed framework is composed of a filtering procedure, which is capable of enhancing the content of the target acoustic event, while reducing the background noise, and an ordinary energy detector dealing with the filtered signal. The improvement can be attributed to two aspects. First, the wavelet packet transform (WPT) is employed for analysis of non-stationary audio signals with a relatively large decomposition scale for precise representation and discrimination between signals. Second, the proposed filter in the wavelet packet domain takes target features into account and is oriented to solve the problem caused by those interwoven features shared by the target and the background noise. Besides the a priori knowledge of the target event, an estimate of the background features is also introduced to the filter design. A complete system for real-time and multi-target detection is presented in this work. Experiments have shown that the proposed method can work with an SNR as low as -15 dB, exhibiting both good reliability and robustness against noise.

The rest of this paper is organized as follows. Section II gives an overview of the system framework and presents a detailed explanation of the proposed filtering method. Section III gives a theoretical analysis of the effectiveness of the filter.

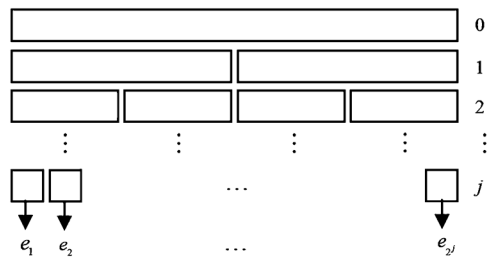


Fig. 2. Wavelet packet tree and its energy spectrum.

Section IV describes the database and protocol used in the experiments and explains how the relevant parameters are chosen. Experimental results and comparisons with existing methods are illustrated in Section V. Finally, Section VI draws conclusions.

II. Detection Algorithm

The proposed algorithm processes a continuous audio stream with a fixed-size sliding window for frame-by-frame detection of a target event, as shown in Fig. 1. The outcomes of overlapped frames are integrated to generate a robust detection result. The dashed box to process each frame comprises a filtering procedure and an energy detector. First, the WPT on samples of the target event class is carried out, and their average forms a target template for configuring the filter. During the online detection, an input frame is decomposed in the time-frequency domain by the WPT as the filtering procedure's input. Next, the filter formed by the target template, as well as the WPT features of the current input, is applied to yield the filtering procedure's output. Finally, an energy detector is employed to measure the energy difference between the filtered output signal and the input signal, i.e., the energy increase after the filtering. Notice that Fig. 1 only shows the detection of a single type of acoustic event. An easy extension can be made for multi-target detection through a multi-channel structure, which will be discussed in Section V-B.

A. Time-Frequency Representation of Signals

Time-frequency analysis shows how the energy of a signal is distributed on the two-dimensional time-frequency plane and is intended for non-stationary signals. There exist many tools for generating a time-frequency representation of a signal. This paper adopts the WPT for acoustic signal analysis, and the results derived in the paper can be easily extended to the cases with other time-frequency analysis tools.

The motivation to use the WPT other than the simpler short-time Fourier transform (STFT) is that the WPT offers a better

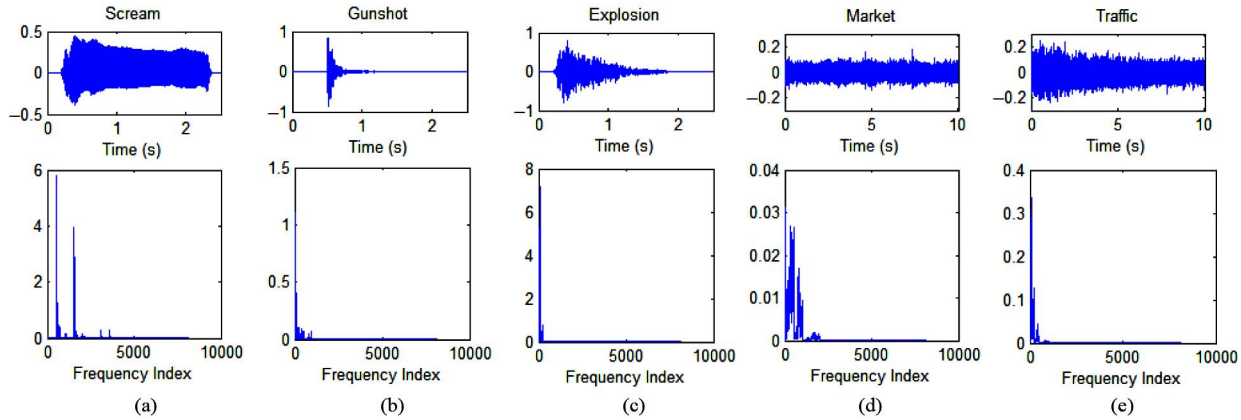


Fig. 3. Raw signals (top) and wavelet packet energy spectrums (bottom) of acoustic events and noises when $j = 13$. (a) Scream. (b) Gunshot. (c) Explosion. (d) Market noise. (e) Traffic noise.

trade-off between time and frequency resolutions with its variable-sized window and can produce a more accurate time-frequency representation. Nevertheless, the STFT is calculated for consecutive segments with a fixed-size window and thus time resolution is lost within each window. In previous works, researchers have found that Fourier-based methods are perfectly suited to the analysis of narrow band signals but not readily adaptable for many transient signals [12]–[14]. Indeed, the WPT has been proved an effective and promising tool in signal representation for transient and non-stationary signal analysis. It can improve the performance of detection and discrimination, as shown in various experimental results. Fast WPT is implemented by an iteration of the filtering with a low-pass and high-pass filter bank, followed by a down-sampling-by-2 procedure. The signal is thus projected into different partitions of equal-sized frequency bands in the Nyquist frequency domain, resulting in a full wavelet packet tree in Fig. 2.

Separating target signal components from background noise can be achieved by filtering in the time-frequency domain. Wavelet-based techniques have been proved successful for noise reduction, termed wavelet de-noising [15], and have been previously used in image filtering, speech enhancement, and transient detection. Generally, wavelet de-noising is implemented by thresholding wavelet coefficients to remove the noise. However, the selection of the threshold is critical and troublesome, as seen in many relevant studies [16] [17]. Moreover, a thresholding technique may lead to the complete loss of the signal under very low SNR conditions.

This paper introduces a frequency domain filter, subband filter, based on the spectral features of acoustic signals. Let $WP(i,k)$ denote the jth level coefficients of the WPT, where $i = 1, \cdots, N$ is the index of frequency bands with $N = 2^j$, and $k = 1, \cdots, M$ is the index of coefficients in each band. The spectral feature is set to be the averaging pth power of the absolute values of wavelet packet coefficients in each band. In this paper, we set $p = 2$, and it is exactly the wavelet packet energy spectrum, denoted by $\boldsymbol{e} = [e_i]_{i=1,\cdots,N}$, where

$$e_i = \frac{1}{M}\sum_{k=1}^{M} |WP(i,k)|^p, p = 2. \tag{1}$$

Fig. 3 shows the wavelet packet energy spectra of three types of acoustic events and two types of noise. It can be observed that the spectral characteristics of different acoustic events and noises are different. The energy spectra of gunshot and explosion events, as well as traffic noise, lie in the very low frequency region. Market noise is more widely distributed in low frequencies, and the scream event is mainly located in two separate regions. The scale factor or the decomposition level j involves a trade-off between time and frequency resolutions and has a direct impact on the decomposed frequency bands. The selection of a proper scale will be further discussed in Section IV-B.

B. Formulation of the Target Guided Subband Filter

The proposed filter in the frequency domain is based on the fact that different types of acoustic events and background noises have different spectral profiles. Ideally, the filter is desired to have the property of passing those frequency bands that belong to the target acoustic event while removing others that are from noise. For an input frame, let $WP_{\text{in}}(i,k)$ be its jth level wavelet packet coefficients with dimension $N \times M$ and $N = 2^j$, $H(i)$ be the subband filter, and $WP_{\text{out}}(i,k)$ be the output after the filtering procedure, i.e.,

$$WP_{\text{out}}(i,k) = H(i) \cdot WP_{\text{in}}(i,k), \quad i = 1, \cdots, N; k = 1, \cdots, M. \tag{2}$$

Assuming that the target acoustic event has a short duration and comparably few time-varying components, the structure of a preliminary filter based on the target spectral features alone can be given in the following form:

$$H(i) = e_{si}, i = 1, \cdots, N \tag{3a}$$

where e_{si} is the ith element of the wavelet packet energy spectrum of the target acoustic event according to (1). It means each value in the filter vector is an average energy in the ith frequency band of the target signal. Thus, after filtering, frequency bands of the input signal are enhanced in proportion to their contribution to the target signal. For practical use, e_{si} is calculated from a dataset of the target acoustic event class and averaged to form a template of the target. In essence, filter (3a) can be considered as a matched filter that correlates a known target template with an unknown input signal in order to detect the presence of the target in the input.

It can be seen that the filter defined by (3a) works well, especially when the target and the background have comparatively uncorrelated energy spectrum components. However, it is more realistic in most cases that these two feature sets should be interwoven with some shared components. The filter considering only target features in (3a) will amplify these shared features regardless of their sources. Furthermore, if these shared features play a dominant role over other unique features in the target, i.e., in the case of a weak target signal versus a strong similar background interference, false detection may happen. Therefore, background interference, especially those containing common features with the target, should be taken into account with care. An improvement can be made to enhance robustness based on the idea of background suppression, which was proposed by Feng *et al.* for image tracking [18].

Considering that the input signal contains information about the background noise, an improved filter in the subbands is proposed as follows:

$$H(i) = \frac{e_{si}}{e_{ni}}, i = 1, \cdots, N \tag{3b}$$

where e_{ni} is the ith element of the energy spectrum of the input signal, regarded as an energy spectrum estimate of the background noise. It is introduced as the denominator into the previous filter, with the aim of increasing the contribution of significant target features while reducing the influence of background features. Values in the filter vector can then be considered as a series of subband-adaptive gains based on the ratios between the target features and the background features. Frequency bands or features in the target that are larger than that in the noise are enhanced. However, frequency bands that have a similar strength as that in the noise, even though they are significant in the target, are suppressed with smaller gains as they cannot be distinguished from noise. In summary, the proposed filter, to some extent, can be seen as a band-pass filter, and the pass rate of a certain frequency band is determined by the ratio between the target energy and the estimated noise energy within that band. To have a better and sufficient estimate of the noise energy spectrum from the input signal, the length of the input signal is empirically set to be above 10 times the length of the target event to be detected.

According to (3b), the filter is determined by the target template, i.e., the average energy spectrum of the target event class, the energy spectrum of the input signal, and the parameters of the WPT. Some intuitive properties of the proposed filter can be observed:

1) It is a target-guided filter that aims to detect one specific type of acoustic event from noise. Those unique or significant features in the target play a key role in the detection. In fact, this filtering process works as long as the target event has some distinctive features compared to the background, which is true in most cases.
2) The filter is adjusted in real-time to adapt to the varying input signal or background noise, as in the denominator of the filter. Thus, it can be applied to different or time-varying environments without any retraining because of its adaptability to the background noise.
3) The goal of this work is to detect and locate a target event in an audio stream. Unlike conventional filtering or de-noising methods for signal reconstruction, the original signal cannot be recovered from the modified wavelet packet coefficients.

The proposed filter will be further discussed in Section III. It can be proved that the proposed filter has selective signal-boosting capability that can enhance the target content in a noisy background for low SNR conditions. More specifically, it confirms that after filtering, the energy increase at the target is more likely to be greater than that at the noise. Therefore, detection can be accomplished even with a simple energy detector, which can be easily implemented in real-time.

C. Detection Algorithm and Performance Analysis

In order to create filter (3b) for the detection of a target event, a target template needs to be obtained offline. A dataset of the target event class is first captured. The jth level WPT and the corresponding energy spectrum are calculated using (1) for each sound in the dataset of the target. After necessary normalization, the results are averaged to form the target template, denoted by $\boldsymbol{e}_s = [e_{si}]_{i=1,\cdots,2^j}$.

Detection Algorithm for an Input Frame: Input: target template $\boldsymbol{e}_s = [e_{si}]_{i=1,\cdots,2^j}$, input frame $x(n)$, $n = 1, \cdots, L_f$, which is the sampling of a segment of audio signal with temporal duration T_f seconds, and parameters of the WPT.

Step 1: Calculate the WPT for $x(n)$ to the jth level, the same scale as that of the target. The result is denoted by $WP_{\text{in}}(i,k)$ with dimension $N \times M$ and the corresponding energy spectrum is $\boldsymbol{e}_n = [e_{ni}]_{i=1,\cdots,2^j}$.

Step 2: Generate the filter and conduct the filtering procedure as follows :

$$WP_{\text{out}}(i,k) = H(i) \cdot WP_{\text{in}}(i,k) \text{ with } H(i) = \frac{e_{si}}{e_{ni}}$$

.

Step 3: Calculate the energy increase sequence $\Delta E(n)$ for the current frame[1]. Let $y(n)$ be the reconstructed signal from $WP_{\text{out}}(i,k)$ through the inverse WPT, and both $x(n)$ and $y(n)$ are normalized in energy as follows:

$$E_{\text{in}}(n) = \frac{|x(n)|^2}{\sum_{k=1}^{L_f} |x(k)|^2}, n = 1, \cdots, L_f$$

$$E_{\text{out}}(n) = \frac{|y(n)|^2}{\sum_{k=1}^{L_f} |y(k)|^2}, n = 1, \cdots, L_f$$

For smoothness of the measured energy increase, $\Delta E(n)$ is summed up in an energy accumulating window with length L_e:

$$\Delta E(n) = \sum_{k=n}^{n+L_e-1} E_{\text{out}}(k) - \sum_{k=n}^{n+L_e-1} E_{\text{in}}(k), n = 1, \cdots, L_f - L_e + 1$$

[1]Energy values can be calculated either in the frequency domain or back in the time domain since they are equivalent by Parseval's theorem.

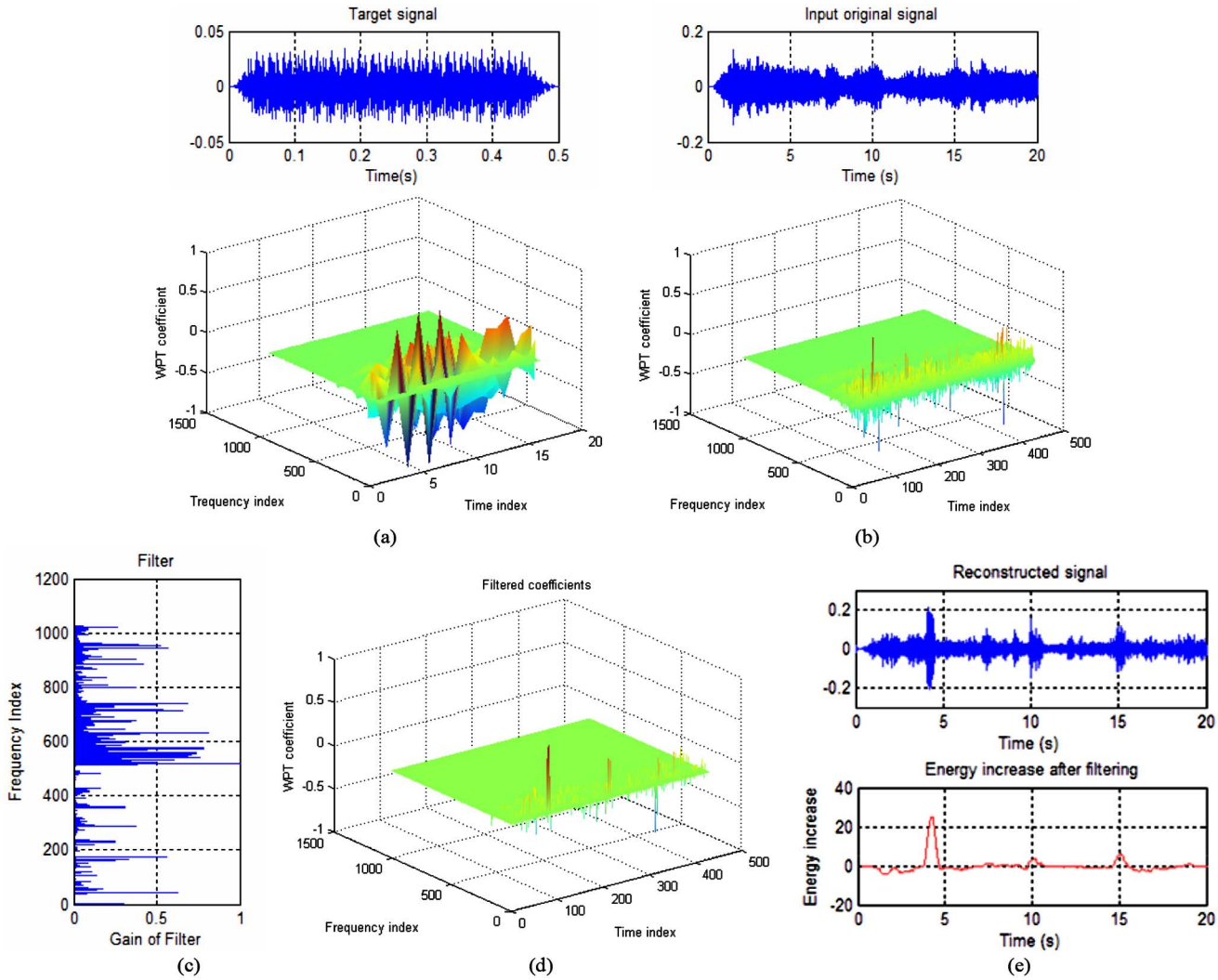


Fig. 4. Detection process of a siren mixed with market noise at 4 s and -10 dB. (a) The target signal and its WPT coefficients. (b) The input signal and its WPT coefficients. (c) The subband filter. (d) The filtered WPT coefficients. (e) The reconstructed signal and the energy increase curve.

Step 4: Detection is made by a threshold, yielding a decision sequence defined by the following:

$$d(n) = \begin{cases} 1, & \text{if } \Delta E(n) > threshold \\ 0, & \text{else} \end{cases}.$$

Fig. 4 shows the whole detection process for an input frame. The test signal is a record of market noise with a -10 dB siren inserted at 4 s. With such a low SNR, it is hard to spot the siren from the temporal wave in Fig. 4(b). However, after filtering, the content of the target is greatly enhanced and the energy increase curve in Fig. 4(e) exhibits a remarkable peak at the moment when the event occurs.

Detection strategy for a continuous audio stream: Detection for a continuous audio stream is an extension of the single frame algorithm, as shown in Fig. 5.

Step 1: Extract overlapped frames of length T_f from the input continuous audio stream with an increment T_i by which T_f is divisible.

Step 2: Conduct the detection process for each frame according to the algorithm mentioned above, resulting in an outcome of energy increase sequence.

Step 3: Outcomes of the overlapped frames are averaged to generate a final robust result to achieve online detection.

In Fig. 5, the test signal is an explosion event merged into subway noise, with SNR being -15 dB, and $T_f = 20$ s, $T_i = 5$ s.

A comparison of performance between filters defined by (3a) and (3b) is illustrated in Fig. 6. The test signal is an explosion event merged into traffic noise. The target and the noise share many similar frequencies and are easily confused, as shown in Fig. 3. Due to the background interference, the energy curve resulting from filter (3a) shows many false peaks, with their levels even higher than that of the real target event, as shown in the left of Fig. 6, which leads to false detections when the SNR goes down to -20 dB. However, by using filter (3b), the interference signals are significantly attenuated, and the target event is highlighted with a remarkable energy increase, as shown in the right of Fig. 6. Filter (3b) outperforms (3a), with a higher energy increase at the target, fewer false peaks, and a better performance, even under very low SNR conditions. The results verify the importance of background suppression, especially when dealing

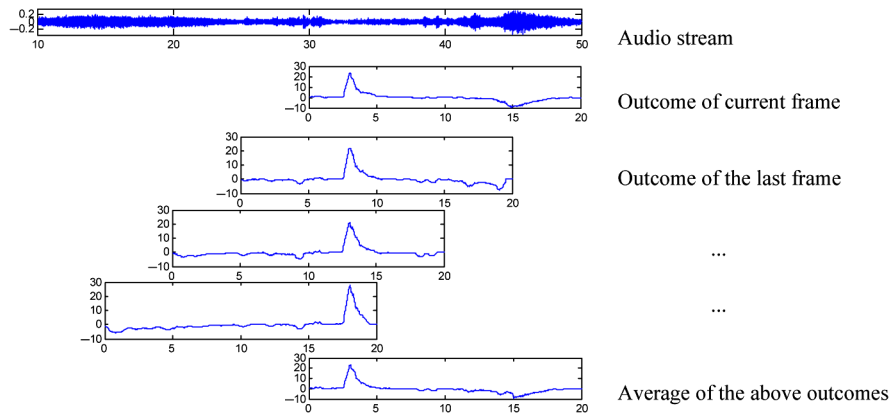


Fig. 5. Process of data fusion of overlapped frames by averaging.

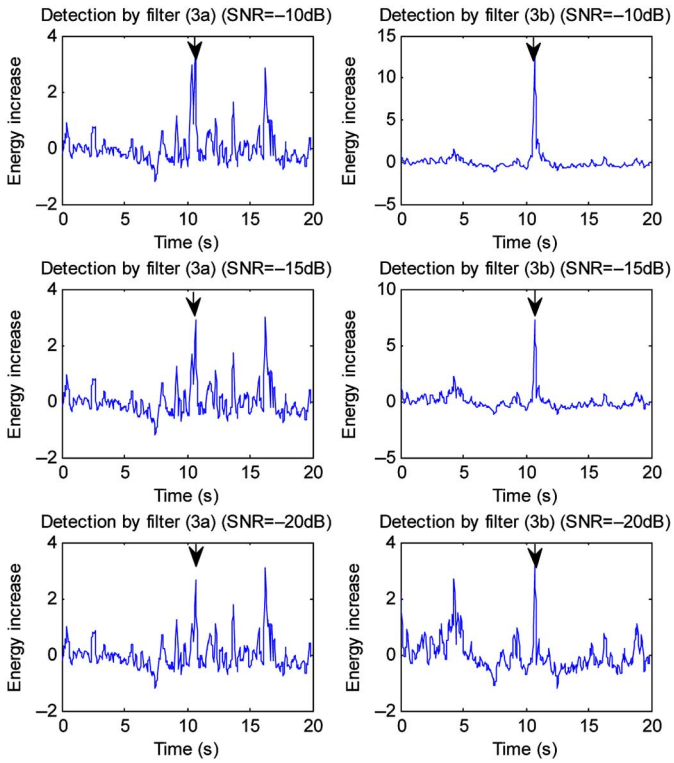


Fig. 6. Performance comparison between filters defined by (3a) (left) and (3b) (right) for detecting an explosion embedded in traffic noise at 10 s under different SNRs: −10 dB (top), −15 dB (middle), and −20 dB (bottom).

with cases having many interwoven features between the background and the target.

III. Theoretical Analysis of the Proposed Filter

In this section, the effectiveness of the proposed filter is discussed theoretically. It will be proved below that the filter in (3b) is capable of enhancing target content embedded in background noise and can offer a high probability of correct detection, even simply thresholding the energy increase.

Consider an input audio frame that is a mix of a target acoustic event and background noise and let $\boldsymbol{A} = [\boldsymbol{a}_{ik}]_{N\times M} \in \mathbb{R}^{N\times M}$ be its time-frequency representation, where N and M stand for dimensions of the frequency and time domains, respectively. In this paper, $\boldsymbol{A}$ is defined by the wavelet packet decomposition (see Section II-A). Time-frequency models and their energy matrices for the target signal and pure noise in the input are introduced as target model:

$$\begin{bmatrix} \hat{\boldsymbol{m}}_1(1) & \cdots & \hat{\boldsymbol{m}}_1(L) \\ \vdots & \ddots & \vdots \\ \hat{\boldsymbol{m}}_N(1) & \cdots & \hat{\boldsymbol{m}}_N(L) \end{bmatrix} \in \mathbb{R}^{N\times L} \text{ with energy } \begin{bmatrix} \hat{\boldsymbol{m}}_1^2(1) & \cdots & \hat{\boldsymbol{m}}_1^2(L) \\ \vdots & \ddots & \vdots \\ \hat{\boldsymbol{m}}_N^2(1) & \cdots & \hat{\boldsymbol{m}}_N^2(L) \end{bmatrix},$$

and noise model:

$$\begin{bmatrix} \boldsymbol{b}_1(1) & \cdots & \boldsymbol{b}_1(M) \\ \vdots & \ddots & \vdots \\ \boldsymbol{b}_N(1) & \cdots & \boldsymbol{b}_N(M) \end{bmatrix} \in \mathbb{R}^{N\times M} \text{ with energy } \begin{bmatrix} \boldsymbol{b}_1^2(1) & \cdots & \boldsymbol{b}_1^2(M) \\ \vdots & \ddots & \vdots \\ \boldsymbol{b}_N^2(1) & \cdots & \boldsymbol{b}_N^2(M) \end{bmatrix}.$$

Every row in the two energy matrices is assumed to be a sequence of independent and identically distributed (i.i.d.) random variables, denoted by $\hat{\boldsymbol{m}}_i^2(\cdot)$ and $\boldsymbol{b}_i^2(\cdot)$, respectively. Moreover, $\hat{\boldsymbol{m}}_i^2(\cdot)$ and $\boldsymbol{b}_i^2(\cdot)$ are mutually independent. Random sequences between different rows in either of the matrices are also mutually independent. Based on these assumptions, it follows that both the target signal and the noise are stationary. It must be pointed out that this stationary requirement for noise is only restricted to the current input frame rather than the whole processing audio stream. Their expectations are denoted as $E\{\hat{\boldsymbol{m}}_i^2(k)\} = \overline{\hat{m}_i^2}$ and $E\{\boldsymbol{b}_i^2(k)\} = \overline{b_i^2}$, respectively.

Suppose that a target event happens at time s. The target is inserted into the background noise with a certain SNR, which affects only L columns (called target zone below) in $\boldsymbol{A}$ with $L << M$. $\boldsymbol{A}$ can then be expressed as,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{b}_1(1)\cdots\boldsymbol{b}_1(s-1) & \boldsymbol{c}_1(1)\cdots\boldsymbol{c}_1(L) & \boldsymbol{b}_1(s+L)\cdots\boldsymbol{b}_1(M) \\ \vdots \quad\ddots\quad \vdots & \vdots \quad\ddots\quad \vdots & \vdots \quad\ddots\quad \vdots \\ \boldsymbol{b}_N(1)\cdots\boldsymbol{b}_N(s-1) & \boldsymbol{c}_N(1)\cdots\boldsymbol{c}_N(L) & \boldsymbol{b}_N(s+L)\cdots\boldsymbol{b}_N(M) \end{bmatrix}$$

where the random variables from the sth to $(s+L-1)$th column in the pure noise model are replaced by $\boldsymbol{c}_i(\cdot)$and $1 \le s \le M - L + 1$. The corresponding energy $\boldsymbol{c}_i^2(\cdot)$ is a mix of energies of the target and noise with the expectation that $E\{\boldsymbol{c}_i^2(k)\} = \overline{c_i^2}$.

In order to obtain a proper representation for $\boldsymbol{c}_i^2(\cdot)$, energy normalization is carried out on the target model. Assume that $\sum_{i=1}^{N} \overline{b_i^2} \approx \sum_{i=1}^{N} \overline{c_i^2}$ for low SNR scenarios, which is the focus of this paper. Then the noise energy can be estimated by $\sum_{i=1}^{N} \overline{b_i^2} \approx 1/M \sum_{i,k} a_{ik}^2$, where a_{ik} is a sample of the

random variable $\boldsymbol{a}_{ik}$ in $\boldsymbol{A}$. The target model is normalized by the following equations:

$$\boldsymbol{m}_i(k) = \sqrt{\phi}\hat{\boldsymbol{m}}_i(k), i = 1, \cdots, N; k = 1, \cdots, L \tag{4}$$

where $\phi = \sum_{i=1}^{N} \overline{b_i^2} / \sum_{i=1}^{N} \overline{\hat{m}_i^2}$. Thus $\overline{m_i^2} = E\{\mathbf{m}_i^2(k)\} = \phi\overline{\hat{m}_i^2}$ and $\sum_{i=1}^{N} \overline{m_i^2} = \sum_{i=1}^{N} \overline{b_i^2}$. Note that $\overline{m_i^2}$ refers to the average energy of the ith frequency band and forms the template of the target, denoted by $\boldsymbol{m} = [\overline{m_i^2}]_{i=\underline{1,\cdots,N}}$. Based on the fact that $\overline{c_i^2}$ is likely to be larger for a larger $\overline{m_i^2}$, one reasonable form for $\boldsymbol{c}_i^2(\cdot)$ can be expressed as,

$$\boldsymbol{c}_i^2(\cdot) = \alpha\boldsymbol{m}_i^2(\cdot) + (1-\alpha)\boldsymbol{b}_i^2(\cdot), 0 < \alpha \leq 1 \tag{5}$$

where α represents the effect of signal mixing. It follows that $\boldsymbol{c}_i^2(\cdot)$ is also an i.i.d. random sequence. Its mathematical expectation satisfies that $\overline{c_i^2} = \alpha\overline{m_i^2} + (1-\alpha)\overline{b_i^2}$, and it is easy to prove that $\sum_{i=1}^{N} \overline{c_i^2} = \sum_{i=1}^{N} \overline{b_i^2}$.

Now the proposed filter in (3b) can be formulated in terms of the target template $\boldsymbol{m}$ and $\boldsymbol{A}$. Let $c_i(k)$ and $b_i(k)$ denote samples of the random variables in $\boldsymbol{A}$. Then we obtain

$$\begin{aligned} H_i &= \frac{M\overline{m_i^2}}{\sum_{k=1}^{M} a_{ik}^2} \\ &= \frac{M\overline{m_i^2}}{\sum_{l\in\{1,\cdots,s-1\}\cup\{s+L,\cdots,M\}} b_i^2(l) + \sum_{l\in\{1,\cdots,L\}} c_i^2(l)} \\ &\approx \frac{M\overline{m_i^2}}{M\overline{b_i^2} + L\alpha(\overline{m_i^2} - \overline{b_i^2})}. \end{aligned} \tag{6}$$

The main properties of the filter can be presented in the following three propositions.

Proposition 1: For the subband filter in (3b), if $M > L$, the ith band gain satisfies that $H_i \begin{cases} > 1, & \overline{m_i^2} > \overline{b_i^2} \\ = 1, & \overline{m_i^2} = \overline{b_i^2} \\ < 1, & \overline{m_i^2} < \overline{b_i^2} \end{cases}$.

Proposition 1 states that the proposed subband filter can adaptively change its gain for any frequency band i according to the relative energies between the target and the noise in the band. It will amplify the signals in the band with $H_i > 1$ if the template energy is stronger than the noise energy but attenuate the signals with $H_i < 1$ otherwise.

In order to further evaluate the performance of the filter for target content enhancement, the difference between the energy increases at the target and at the noise needs to be analyzed. The energy increase between the filtered output and the original input calculated in the time-frequency domain referring to a single row, i.e., a frequency band, is

$$\boldsymbol{E}_i(l) = \begin{cases} (H_i\boldsymbol{c}_i(l-s+1))^2 - \boldsymbol{c}_i^2(l-s+1), & l\in\{s,\cdots,s+L-1\} \\ (H_i\boldsymbol{b}_i(l))^2 - \boldsymbol{b}_i^2(l), & l\in\{1,\cdots,s-1\}\cup\{s+L,\cdots,M\}. \end{cases} \tag{7}$$

Considering a single column k within the target zone, the difference between its energy increase and that of any other noise column l is

$$\begin{aligned} \Delta\boldsymbol{E}_i(k,l) &= \boldsymbol{E}_i(k) - \boldsymbol{E}_i(l) \\ &= (H_i^2 - 1)\boldsymbol{c}_i^2(k) - (H_i^2 - 1)\boldsymbol{b}_i^2(l) \\ &= (H_i^2 - 1)(\alpha(\boldsymbol{m}_i^2(k) - \boldsymbol{b}_i^2(k+s-1)) \\ &\quad + (\boldsymbol{b}_i^2(k+s-1) - \boldsymbol{b}_i^2(l))) \end{aligned} \tag{8}$$

where $k \in \{1,\cdots,L\}$ and $l \in \{1,\cdots,s-1\} \cup \{s+L,\cdots,M\}$. Thus, the total difference over all frequency bands is summarized as

$$\Delta\boldsymbol{E}(k,l) = \sum_{i=1}^{N} \Delta\boldsymbol{E}_i(k,l). \tag{9}$$

It is expected that the difference is positive and as large as possible in order to increase the probability of correct detection for the energy based target detection in Step 4.

Proposition 2: Through the subband filter in (3b), the target signal will gain more energy increase than the noise, i.e., $E\{\Delta\boldsymbol{E}(k,l)\} \geq 0$, and it equals 0 if and only if $\overline{m_i^2} = \overline{b_i^2} \forall i$.

Proposition 2 shows that, for any frequency band i, the mathematical expectation of the energy difference $\Delta\boldsymbol{E}_i(k,l)$ is always positive, as long as there exists an energy difference between the target model and the noise model, i.e., $\overline{m_i^2} \neq \overline{b_i^2}$. It confirms that the proposed filter has selective signal boosting capability that can always enhance the target content in a noisy background for whatever $\overline{m_i^2} > \overline{b_i^2}$ or $\overline{m_i^2} < \overline{b_i^2}$. Thus the proposed filter is suitable for low SNR situations.

The next proposition further compares the probability $P\{\Delta\boldsymbol{E}(k,l) > 0\}$ with $P\{\Delta\boldsymbol{E}_i(k,l) > 0\}$ of individual i, which can illustrate how the scale of the wavelet packet decomposition affects the probability of correct detection.

Proposition 3: If $\boldsymbol{m}_i^2(k_1) - \boldsymbol{b}_i^2(k_2)$ and $\boldsymbol{b}_i^2(k_3) - \boldsymbol{b}_i^2(k_4)$, $k_1 \in \{1,\cdots,L\}$, $k_2, k_3, k_4 \in \{1,\cdots,M\}$, $k_3 \neq k_4$, for any frequency band $i \in \{1,\cdots,N\}$, are normally distributed, the probability of correct detection $P\{\Delta\boldsymbol{E}(k,l) > 0\} = \Phi(g)$ increases from $P\{\Delta\boldsymbol{E}_i(k,l) > 0\} = \Phi(g_i)$ for a single band i with $g_i \leq g \leq \sqrt{N}g_i$, where $\Phi(\cdot)$ is the distribution function of the standard normal variable.

From proposition 3, g is $1 \sim \sqrt{N}$ times the amount of g_i. It means that the probability of the energy difference being positive, which is calculated over all frequency bands, can be improved remarkably compared with the probability of that in a single frequency band. It provides a way to improve the probability of correct detection in applications, especially for low SNR situations. A large N is likely to achieve a high probability of correct detection, which requires large scale wavelet packet decomposition.

The proofs of the three propositions are given in the Appendix.

IV. Experimental Set-Up

A. Database and Metric

The database used in the experiments includes three types of acoustic events: screams (184 sounds), gunshots (196 sounds),

and explosions (153 sounds), and various noises: white, restaurant, market, traffic, and subway, each with a duration of about 4 min. Sound samples were collected from various public sound effects libraries: BBC Sound Effects Library, Sound Ideals Series 6000, Sound Ideas: the art of Foley and Best Service Studio Box Sound Effects. All sounds were sampled at 20.05 kHz with a 16-bit resolution. The composition of the database (including the acoustic events and the subway noise) was intended to be similar to the one used in [1].

In order to produce the target template which is needed for generating the filter, the dataset of each acoustic event class was randomly divided into 75% for template training and 25% for testing. The collection of the sounds took into account their diversity as well as the similarity between the samples. The experimental results can also verify the sensitivity of the filter to the target dataset for template training.

To validate the proposed algorithm, we built a simulation dataset, and each file, about 4 min in length, was generated by inserting an acoustic event from the target test dataset into a specific type of noise at several random moments with a certain SNR. The test SNR went from 5 dB to -15 dB, with a 5 dB step.

The performance of the proposed system was measured using the detection error tradeoff (DET) curve, which takes into account both the MDR and the FDR. A missed detection means non-detection of a real event, and a false detection means detection of a non-existing event. These two error rates are closely tied to the sensitivity of the system, with reference to the threshold of the energy detector in the proposed framework. An overall detection performance is characterized by the EER. It is defined as the value of MDR or FDR when these two rates are equal in the DET curve. The threshold is manually changed to get different pairs of MDRs and FDRs necessary to plot DET curves.

B. Choosing the Parameters

In the implementation of the proposed detection algorithm, the wavelet basis and the decomposition level j of the WPT are two major parameters that affect system performance.

Selection of the wavelet basis: Wavelet analysis offers the flexibility of using a number of basis functions, and the most famous wavelet families include Haar, Morlet, Marr, Daubechies, Coiflet, etc. The choice of the wavelet basis could be critical in the quality of the signal description due to their different properties. For specific detection problem, the criterion used in this paper was to synthetically consider the mathematical properties of the basis functions and their experimental performance.

This work applied the Coiflet of order 5, which exhibits good symmetry and orthogonality. Extensive tests were carried out on various basis functions such as db10, coif5, etc. Their performances were evaluated by the average energy increase at the target because of the fact that the larger the energy increase is, the higher the probability of correct detection is. Results in Table I show that coif5 produces the largest energy increase and outperforms the other basis functions.

Selection of the Decomposition Level j : The decomposition level or the scale factor j is another key parameter that directly determines the number of subbands after decomposition. As mentioned in Section II-A, a larger scale means a finer partition of frequencies and thus offers better discrimination between the target event and noise, which is indeed necessary for low SNR situations. Moreover, from the theoretical analysis in Section III, Proposition 3 verifies that a large scale helps improve the probability of correct detection. However, it should not be too high due to the consequent heavy computational load. There is also an upper limit due to the finite length of the signal.

TABLE I
AVERAGE ENERGY INCREASE VALUES OF DETECTION WITH DIFFERENT BASIS FUNCTIONS

Test	db5	db10	coif3	coif5	sym8
Explosion+Market	23.74	30.18	27.30	**38.18**	29.51
Scream+Market	23.18	21.62	21.99	**28.88**	23.38
Gunshot+Market	14.22	17.11	15.51	**23.97**	18.29

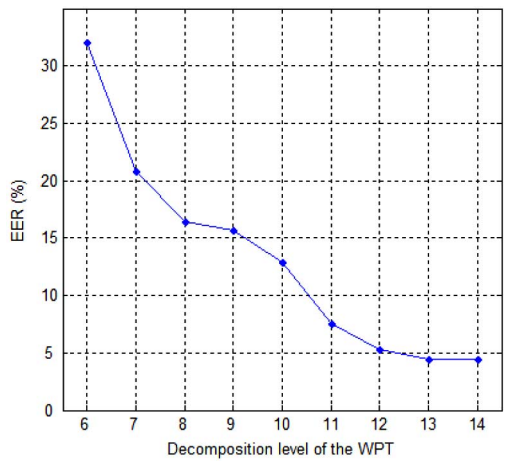


Fig. 7. EER as a function of the decomposition level of the WPT for detection at -15 dB.

As the focus of this work is to improve the detection performance with an adverse background, we evaluated the EERs for different levels of j in order to choose the optimal one. A series of experiments have been done for target detection with respect to different decomposition levels varying from 6 to 14. Results in Fig. 7 show that the EER of detection falls significantly as j increases from 6 to 11, and when $j > 11$, the improvement is no longer notable. In the experiments we set $j = 13$.

For the other parameters, we used $T_f = 20$ s and $T_i = 1$ s for the frame-extracting window and the temporal duration T_e of the energy accumulating window was 0.5 s. To deal with the heavy computation load of the 13-level WPT, the wavelet analysis was hardware accelerated by mapping it onto a graphic processing unit (GPU). The simulation was conducted on a PC with a 2.53 GHz Intel Xeon CPU and C++ implementation combined with the NVIDIA Tesla C1060 computing processor. It was verified that the whole detection algorithm takes about 0.225 s to process one frame, and thus is able to satisfy the demand of real-time applications.

TABLE II
EQUAL ERROR RATES (%) OF THE PROPOSED METHOD UNDER DIFFERENT NOISES WITH RESPECT TO SINGLE TARGET DETECTION

Type	SNR(dB)	White	Restaurant	Market	Traffic	Subway
Screams	5	0	0	0	0	5.88
	0	0	0	0	0	7.69
	–5	0	4.93	7.11	4.93	9.43
	–10	6.84	15.67	22.29	11.35	10.47
	–15	10.8	39.66	36.11	12.5	17.95
Gunshots	5	0	0	0	0	0
	0	0	0	0	0	0
	–5	0	0	0	0.12	1.69
	–10	0	0	0.21	1.46	4.38
	–15	0.63	0.51	2.61	2.76	7.95
Explosions	5	0	0	0	0	0
	0	0	0	0	0	0
	–5	0	0	0	1.24	1.86
	–10	0	0	1.43	5.66	5.03
	–15	1.41	2.04	4.29	12.53	12.01

TABLE III
OVERALL EQUAL ERROR RATES OF THE PROPOSED METHOD COMPARED WITH THE MEDIAN FILTER BASED METHOD WITH RESPECT TO SINGLE TARGET DETECTION

SNR(dB)	EER(%) of proposed method		EER(%) of median filter based method [5]	
	White	Market	White	Market
5	**0**	**0**	0	7.51
0	**0**	**0**	0.35	23.46
–5	**0**	**2.37**	5.86	53.72
–10	**0**	**7.91**	18.21	-
–15	**5.01**	**13.9**	40.57	-

V. EXPERIMENTS

A. First Experiment: Single Target Detection

The first experiment was carried out by following the single target detection procedure in Fig. 1. The goal was to verify the impact of noisy environments on detection performance. Experiments were done individually for scream detection, gunshot detection, and explosion detection. Several types of noise (white, restaurant, market, traffic, and subway) were tested. Table II gives the EERs with respect to the three events mixed with different noises at different SNRs. We have also compared the performances of our method and the ordinary energy detector based on a median filter proposed by Dufaux *et al.* [5]. Results in Table III show a significant improvement with our method, especially at low SNRs.

In terms of performance with different background noises, the best results were achieved with white noise, apparently because of its excellent stationary property and less embedded interference. The detectable SNR range under white noise was brought down to as low as -15 dB, an outstanding result that demonstrates the filter's capability in reducing the noise at low SNRs. As expected, the results degrade a little with real environmental noises. However, a nearly errorless detection above 0 dB and an average EER of 13.3% at -15 dB with various noise sources were achieved. This confirms that the proposed method exhibits robustness and is capable of practical applications. Table II shows the best detected event is the gunshot compared to the other two events, and an average EER less than 5% above -15 dB was achieved. This may owe to the very concentrated distribution of its template in the low frequencies, which will be beneficial for the subband filtering procedure. However, the scream event was worst detected due to the large diversity of people's scream sounds.

Table II also shows that the detection performance differs with respect to different combinations of acoustic events and noises, which can be possibly explained by the different profiles of the spectra of these signals (see Fig. 3). The traffic and subway noises are mainly concentrated in very low frequencies, which are similar to the gunshot and explosion events but distinct from the scream event. In the other cases, the market and restaurant noises contain many vocal sounds, and hence are more of a hindrance to scream detection. For example, talking or laughing are often detected as screams when the SNR goes down below a critical point. Thus, results for scream detection with market and restaurant noises are worse than those under traffic and subway noises, while it is the reverse for gunshot and explosion detection.

B. Second Experiment: Multi-Target Detection

In previous sections, a filtering-based algorithm was developed, which was limited to single target detection. For more practical and complex applications, we extended it to a more general framework for multi-target detection as shown in Fig. 8. In this work, we concentrated on detecting screams, gunshots, and explosions in a subway background. The multi-target detection system comprised three detection channels that run in parallel for each of the three events. The target detector in each channel is described in Fig. 1. To complete detection, the data fusion of overlapped frames is implemented on the outputs of three channels, resulting in three energy increase curves for three events, respectively. The final decision is made according to an intuitive strategy: a target event is detected if at least one channel's output exceeds the threshold, and its type is determined by the channel that presents the maximum energy increase. It is also convenient to add other events of interest to the system.

For the proposed multi-channel detection system, false detections in the three channels will be accumulated. Furthermore, due to the high correlation between events, e.g., gunshots and explosions, which share many similar frequencies (see Fig. 3), misclassification may be the main concerning factor for performance evaluation. It is worth mentioning that gunshots under

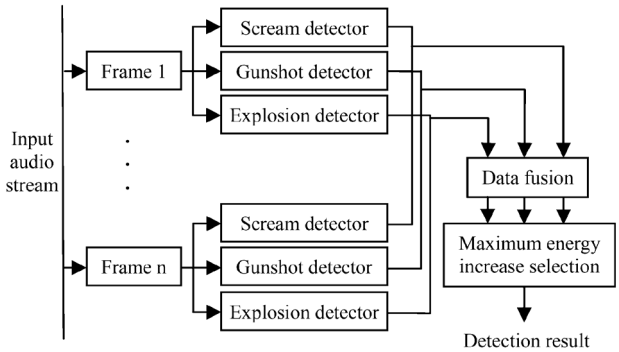


Fig. 8. Framework of the multi-target detection system.

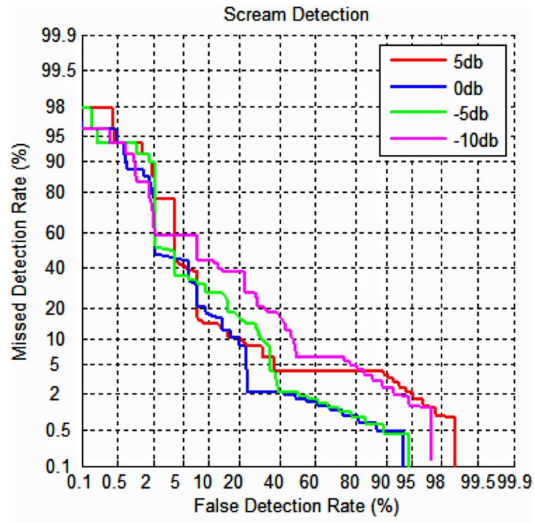
subway noise misclassified as explosions was previously reported by Ntalampiras *et al.* [1], and the EER for gunshot detection remained high (about 25%) for an SNR varying from 15 dB to −5 dB. In order to reduce acoustic confusion between different event classes and finally improve performance, other features were taken into account. Considering that the three events have different durations, gunshots have a mean duration of about 0.5 s while an explosion can last 3 s or longer. Therefore, energy accumulating windows with different duration T_e were applied for computing energies, that is, 0.5 s for gunshot detection and 2.5 s for explosion detection.

Experiments were done for the three target detection under subway noise in our database. Table IV gives the EERs of the proposed method compared with the HMM-based method [1]. The data in the last column of Table IV is directly derived from [1, Table VI] on the basis that we use a similar database and experimental protocol. Results show that the performance degrades greatly compared to single-target detection due to confusion between different events. An average EER for three events at −10 dB is 21.45%. More precisely, explosions are detected with EER of 14.29%, gunshots with 23.08%, and screams with 27.08%. In particular, the EERs for gunshot detection are brought down notably. Fig. 9 depicts the DET curves with respect to the detection of each target event class at different SNRs.

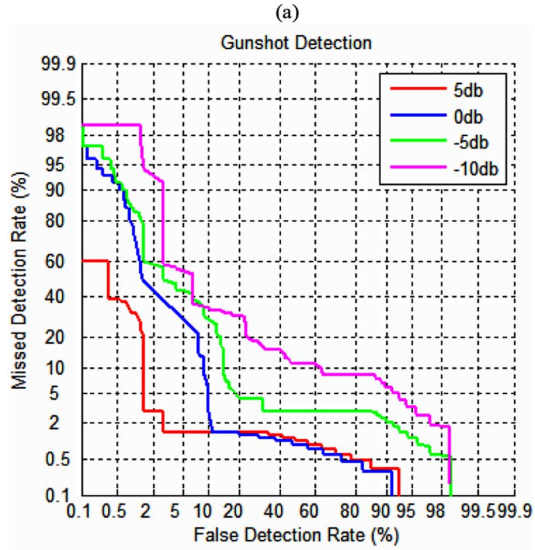
VI. Conclusion

In this paper, an effective filtering approach using wavelet packets is proposed for AED under low SNR conditions. Taking advantage of the time-frequency representation by the WPT, acoustic signals can be filtered in subbands to separate target components from noise. The proposed subband filter considers that different acoustic events and noises show different spectral characteristics. Background noise can be effectively suppressed by simultaneously taking into account the target spectrum and an estimate of the noise spectrum. In fact, this filter can be considered as a band-pass filter that can automatically pass frequency bands that are more significant in the target than in the noise. It is proved that the filtering method is capable of enhancing the target content, while suppressing the background noise under low SNR conditions. It is also shown that a larger decomposition level of the WPT helps to improve the probability of correct detection.

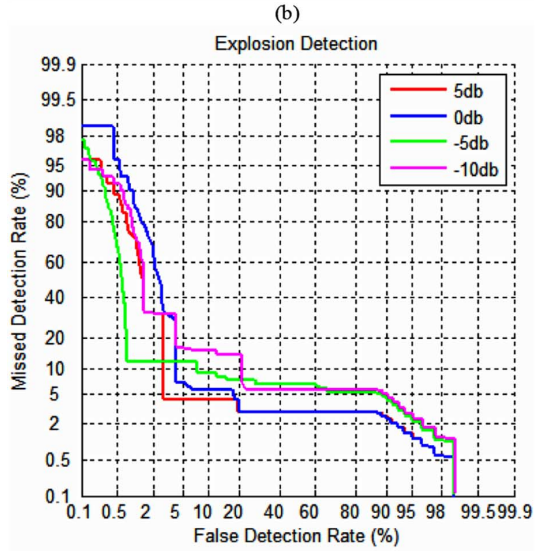
Two series of experiments have been done on a large dataset for single-target and multi-target detection. For single-target detection, the detectable SNR can be brought down to −15 dB with various background noises. This superb performance



(a)



(b)



(c)

Fig. 9. DET curves of three target events under subway noise at different SNRs with respect to multi-target detection. (a) Scream detection. (b) Gunshot detection. (c) Explosion detection.

TABLE IV
EQUAL ERROR RATES OF THE PROPOSED METHOD COMPARED WITH THE HMM BASED METHOD REGARDING TO MULTI-TARGET DETECTION

Type	SNR(dB)	EER(%) of proposed method	EER(%) of HMM based method [1]
Screams	5	**13.52**	16.50
	0	**14.29**	21.42
	−5	**18.69**	28.21
	−10	**27.08**	–
Gunshots	5	**2.74**	25.67
	0	**9.09**	26.32
	−5	**14.63**	26.51
	−10	**23.08**	–
Explosions	5	**4.22**	7.48
	0	**6.22**	8.54
	−5	**9.52**	13.29
	−10	**14.29**	–

demonstrates the effectiveness of the proposed detection solution with the subband filter followed by a simple energy detector. The second experiment extends our method to multi-target detection for screams, gunshots, and explosions. A satisfying overall EER of 14.28% at -5 dB is achieved.

APPENDIX
PROOFS OF PROPOSITIONS

Proposition 1: For the subband filter in (3b), if $M > L$, the ith band gain satisfies that $H_i \begin{cases} > 1, & \overline{m_i^2} > \overline{b_i^2} \\ = 1, & \overline{m_i^2} = \overline{b_i^2} \\ < 1, & \overline{m_i^2} < \overline{b_i^2} \end{cases}$.

Proof: When $\overline{m_i^2} > \overline{b_i^2}$, we have $M\overline{b_i^2} + L\alpha(\overline{m_i^2} - \overline{b_i^2}) < M\overline{m_i^2}$, and hence $H_i > 1$ from (6). Other results can be obtained for the rest two cases in a similar way. □

Proposition 2: Through the subband filter in (3b), the target signal will gain more energy increase than the noise, i.e., $E\{\Delta\mathbf{E}(k,l)\} \geq 0$, and it equals 0 if and only if $\overline{m_i^2} = \overline{b_i^2} \forall i$.

Proof: From (8) and (9), the mathematical expectation is given by

$$E\{\Delta\boldsymbol{E}(k,l)\} = \sum\nolimits_{i=1}^{N} E\{\Delta\boldsymbol{E}_i(k,l)\} \tag{10}$$

where

$$E\{\Delta\boldsymbol{E}_i(k,l)\} = \alpha(H_i^2 - 1)(\overline{m_i^2} - \overline{b_i^2}). \tag{11}$$

From Proposition 1, we have that if $\overline{m_i^2} \neq \overline{b_i^2}$, then $\text{sign}(H_i^2 - 1) = \text{sign}(\overline{m_i^2} - \overline{b_i^2})$, and hence $E\{\Delta\boldsymbol{E}_i(k,l)\} = \alpha(H_i^2 - 1)(\overline{m_i^2} - \overline{b_i^2}) \geq 0$. So $E\{\Delta\mathbf{E}(k,l)\} = \sum_{i=1}^{N} E\{\Delta\boldsymbol{E}_i(k,l)\} \geq 0$ and it equals 0 if and only if $\overline{m_i^2} = \overline{b_i^2} \forall i$. □

Proposition 3: If $\boldsymbol{m}_i^2(k_1) - \boldsymbol{b}_i^2(k_2)$ and $\boldsymbol{b}_i^2(k_3) - \boldsymbol{b}_i^2(k_4)$, $k_1 \in \{1, \cdots, L\}$, $k_2, k_3, k_4 \in \{1, \cdots, M\}$, $k_3 \neq k_4$, for any frequency band $i \in \{1, \cdots, N\}$, are normally distributed, the probability of correct detection $P\{\Delta\boldsymbol{E}(k,l) > 0\} = \Phi(g)$ increases from $P\{\Delta\boldsymbol{E}_i(k,l) > 0\} = \Phi(g_i)$ for a single band i with $g_i \leq g \leq \sqrt{N} g_i$, where $\Phi(\cdot)$ is the distribution function of the standard normal variable.

Proof: Because the period of the target signal is comparatively short, i.e., $L << M$, the filter can be reduced to $H_i \approx \overline{m_i^2}/\overline{b_i^2}$. Equation (8) can then be rewritten as

$$\begin{aligned}\Delta\boldsymbol{E}_i(k,l) = \beta_i(\overline{m_i^2} - \overline{b_i^2})(\alpha(\boldsymbol{m}_i^2(k) - \boldsymbol{b}_i^2(k+s-1)) \\ + (\boldsymbol{b}_i^2(k+s-1) - \boldsymbol{b}_i^2(l)))\end{aligned} \tag{12}$$

where $\beta_i = (\overline{m_i^2} + \overline{b_i^2})/(\overline{b_i^2})^2 > 0$.

Following the conditions of this proposition, we have that $\boldsymbol{m}_i^2(k) - \boldsymbol{b}_i^2(k+s-1)$ and $\boldsymbol{b}_i^2(k+s-1) - \boldsymbol{b}_i^2(l)$ in (12) are normally distributed. It is easy to verify that these two random variables are jointly normal. Thus, $\Delta\boldsymbol{E}_i(k,l)$ is normally distributed based on the fact that for a multivariate normal distribution, any linear combination of its components is normally distributed. Furthermore, $\Delta\boldsymbol{E}(k,l)$ is also normally distributed according to the invariance of linear transform of independent normal random variables. Let $\Delta\boldsymbol{E}_i(k,l) \sim N(\mu_i, \sigma_i^2)$ and $\Delta\boldsymbol{E}(k,l) \sim N(\sum_{i=1}^{N}\mu_i, \sum_{i=1}^{N}\sigma_i^2)$, where $\mu_i = \beta_i(\overline{m_i^2} - \overline{b_i^2})^2 \geq 0$, $\sigma_i^2 = \beta_i^2(\overline{m_i^2} - \overline{b_i^2})^2\hat{\sigma}_i^2$ and $\hat{\sigma}_i^2$ is the variance of the expression in the last brackets of $\Delta\boldsymbol{E}_i(k,l)$ in (12). From the fact that $P\{\boldsymbol{X} > 0\} = \Phi(\mu/\sigma)$ for $\boldsymbol{X} \sim N(\mu, \sigma^2)$, where $\Phi(\cdot)$ is the distribution function of the standard normal variable, the probability increases as the ratio μ/σ increases.

Referring to the energy difference in a single subband $\Delta\boldsymbol{E}_i(k,l)$, we define the ratio $g_i = \mu_i/\sigma_i = |\overline{m_i^2} - \overline{b_i^2}|/\sigma_i$. Then for the energy difference over all subbands $\Delta\boldsymbol{E}(k,l)$, we have

$$g = \frac{\sum_{i=1}^{N}\mu_i}{\sqrt{\sum_{i=1}^{N}\sigma_i^2}} = \frac{\sum_{i=1}^{N}\beta_i(\overline{m_i^2} - \overline{b_i^2})^2}{\sqrt{\sum_{i=1}^{N}\beta_i^2(\overline{m_i^2} - \overline{b_i^2})^2\hat{\sigma}_i^2}}. \tag{13}$$

Define $\boldsymbol{d} = [|\overline{m_i^2} - \overline{b_i^2}|]_{i=1,\cdots,N}$ and $\boldsymbol{B} = diag\{\beta_i\}_{i=1,\cdots,N}$. If $\hat{\sigma}_i = \hat{\sigma} \forall i$, then we obtain

$$g = \frac{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{d}}{\hat{\sigma}\sqrt{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{B}\boldsymbol{d}}} = \frac{(\boldsymbol{B}\boldsymbol{d})^{\mathrm{T}}\boldsymbol{d}}{\hat{\sigma}\,\|\boldsymbol{B}\boldsymbol{d}\|} = \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{d}}{\hat{\sigma}} \tag{14}$$

where $\boldsymbol{u} = \boldsymbol{B}\boldsymbol{d}/\|\boldsymbol{B}\boldsymbol{d}\|$ is a unit vector along $\boldsymbol{B}\boldsymbol{d}$.

1) Considering the best case that maximizes g, we obtain $\boldsymbol{u} \| \boldsymbol{d}$ and it follows that $\beta_i = \beta \forall i$ (for instance, in the case of white noise background). Then $g = \sqrt{\sum_{i=1}^{N}(\overline{m_i^2} - \overline{b_i^2})^2}/\hat{\sigma}$. If $|\overline{m_i^2} - \overline{b_i^2}| = \mu \forall i$, we finally get $g = \sqrt{N}\mu/\hat{\sigma} = \sqrt{N} g_i$.
2) Considering the worst case $\boldsymbol{u} = [0 \ \cdots \ 0 \ 1 \ 0 \ \cdots \ 0]^{\mathrm{T}}$, where only the ith element is 1, we get $g = |\overline{m_i^2} - \overline{b_i^2}|/\hat{\sigma}_i = g_i$.

Therefore, we can conclude $g_i \leq g \leq \sqrt{N} g_i$. □

ACKNOWLEDGMENT

The authors would like to thank Guang-Mei Yun, Wen-Tao Guo, Xin Cheng, Shuo Liu, Zhen Liu, and Zi-Chen Gao for their assistance and helpful discussions and John Baruch who as a native English speaker read and corrected the text.

REFERENCES

[1] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "A practical system for acoustic surveillance of hazardous situations," *Int. J. Artif. Intell. Tools*, vol. 20, no. 1, pp. 119–137, Feb. 2011.

[2] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J.-F. Serignat, "Information extraction from sound for medical telemonitoring," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 2, pp. 264–274, Apr. 2006.

[3] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recogn. Lett.*, vol. 30, no. 14, pp. 1281–1288, Oct. 2009.

[4] A. Chacon-Rodriguez, P. Julian, L. Castro, P. Alvarado, and N. Hernandez, "Evaluation of gunshot detection algorithms," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 2, pp. 363–373, Feb. 2011.

[5] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini, "Automatic sound detection and recognition for noisy environment," in *Proc. EUSIPCO*, Tampere, Finland, 2000, pp. 1033–1036.

[6] T. Ahmed, M. Uppal, and A. Muhammad, "Improving efficiency and reliability of gunshot detection systems," in *Proc. ICASSP*, 2013, pp. 513–517.

[7] J. Moragues, A. Serrano, L. Vergara, and J. Gosalbez, "Acoustic detection and classification using temporal and frequency multiple energy detector features," in *Proc. ICASSP*, 2011, pp. 1940–1943.

[8] S. Chachada and C.-C. J. Kuo, "Environmental sound recognition: A survey," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, pp. 1–9.

[9] R. Cai, L. Lu, A. Hanjalic, H.-J. Zhang, and L.-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 1026–1039, May 2006.

[10] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proc. ICME*, Amsterdam, The Netherlands, 2005, pp. 1306–1309.

[11] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection in noisy environments," in *Proc. EUSIPCO*, Poznan, Poland, 2007, pp. 1–4.

[12] A. Averbuch, V. A. Zheludev, N. Rabin, and A. Schclar, "Wavelet-based acoustic detection of moving vehicles," *Multidimens. Syst. Signal Process.*, vol. 20, no. 1, pp. 55–80, Mar. 2009.

[13] R. E. Learned and A. S. Wilsky, "A wavelet packet approach to transient signal classification," *Appl. Comput. Harmonic Anal.*, vol. 2, no. 3, pp. 265–278, Jul. 1995.

[14] M. R. Canal, "Comparison of wavelet and short time Fourier transform methods in the analysis of EMG signals," *J. Med. Syst.*, vol. 34, no. 1, pp. 91–94, Feb. 2010.

[15] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[16] Y. Xu, J. B. Weaver, D. M. Healy, Jr, and J. Lu, "Wavelet transform domain filters: A spatially selective noise filtration technique," *IEEE Trans. Image Process.*, vol. 3, no. 6, pp. 747–758, Nov. 1994.

[17] P. Ravier and P.-O. Amblard, "Wavelet packets and de-noising based on higher-order-statistics for transient detection," *Signal Process.*, vol. 81, no. 9, pp. 1909–1926, Sep. 2001.

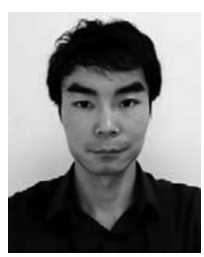
[18] Z. Feng, N. Lu, and P. Jiang, "Posterior probability measure for image matching," *Pattern Recogn.*, vol. 41, no. 7, pp. 2422–2433, Jul. 2008.

Zu-Ren Feng (M'05) received M. Eng. and Ph.D. degrees in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1982 and 1988, respectively. Since 1994, he has been a Professor with the School of Electronic and Information Engineering, Xi'an Jiaotong University. In 1992, he worked as a visiting scholar with INRIA, France, for research on manipulator control with flexible joints and applications of Petri Nets in DEDS. In 1994, he was invited by Kassel University, Germany, for research on mobile service robots. In 2006 and 2007, he worked as a Visiting Professor at the University of Bradford, U.K., for research on multi-agent systems. His research interests include robotics and automation, intelligent information processing, and evolutionary computing based optimization.

Qing Zhou received the B.S. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2012. She is now working toward the M.S. degree in systems engineering in Xi'an Jiaotong University. Her research interests include pattern recognition, signal processing, and acoustic event detection and recognition.

Jun Zhang received the B.S. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2013 He is now working toward the M.S. degree in systems engineering in Xi'an Jiaotong University. His research interests include pattern recognition, signal processing, and brain computer interface.

Ping Jiang is Professor in Computer Science at the University of Hull, UK. He received the B. Eng., M. Eng., and Ph.D. degrees in Information and Control Engineering from Xi'an Jiaotong University, Xi'an, China, in 1985, 1988, and 1992, respectively. He was appointed Lecturer in Department of Electrical Engineering at Tongji University, Shanghai, in 1992 and promoted to Associate Professor in 1994. From 1997 to 2012, he was Professor in Department of Information and Control Engineering at Tongji University. He was Lecturer, Senior Lecturer and Reader in Robotics and Distributed Systems at the University of Bradford from 2003 to 2012. From 1998 to 2000, he was an Alexander von Humboldt Research Fellow in Electrical Engineering at Universitaet Erlangen-Nuernberg, Germany. From 2002 to 2003, he was a Senior Research Fellow in Computing at Glasgow Caledonian University. His research work mainly focuses on signal processing and pattern recognition, intelligent robotics, automation and control, wireless sensor networks, multi-agent and virtual organization.



Xue-Wen Yang received the B.S. degree in electronics engineering from Xi'an Jiaotong University, Xi'an, China, in 2012. From 2010 to 2012, he studied computer science in Ecole Centrale Marseilles (ECM) in France for a double master degree program and received his M.S. degree in computer science from ECM In 2014. He is now working toward the M.S. degree in systems engineering in Xi'an Jiaotong University. His research interests include motion control, manipulation in robotics, communication, and intelligent information processing.