PAPER

An Adaptive Wavelet-Based Denoising Algorithm for Enhancing Speech in Non-stationary Noise Environment

Kun-Ching WANG†, Member

SUMMARY  Traditional wavelet-based speech enhancement algorithms are ineffective in the presence of highly non-stationary noise because of the difficulties in the accurate estimation of the local noise spectrum. In this paper, a simple method of noise estimation employing the use of a voice activity detector is proposed. We can improve the output of a wavelet-based speech enhancement algorithm in the presence of random noise bursts according to the results of VAD decision. The noisy speech is first preprocessed using bark-scale wavelet packet decomposition (BSWPD) to convert a noisy signal into wavelet coefficients (WCs). It is found that the VAD using bark-scale spectral entropy, called as BS-Entropy, parameter is superior to other energy-based approach especially in variable noise-level. The wavelet coefficient threshold (WCT) of each subband is then temporally adjusted according to the result of VAD approach. In a speech-dominated frame, the speech is categorized into either a voiced frame or an unvoiced frame. A voiced frame possesses a strong tone-like spectrum in lower subbands, so that the WCs of lower-band must be reserved. On the contrary, the WCT tends to increase in lower-band if the frame is categorized as unvoiced. In a noise-dominated frame, the background noise can be almost completely removed by increasing the WCT. The objective and subjective experimental results are then used to evaluate the proposed system. The experiments show that this algorithm is valid on various noise conditions, especially for color noise and non-stationary noise conditions.

key words: speech enhancement, bark-scale, time-frequency adaptation, spectral entropy, voice-activity detector

1. Introduction

Automatic speech processing systems are employed more and more often in new applications in a variety of real environments. However, in many practical situations, they must contend with high ambient noise levels, and their performance degrades drastically. Hence, speech enhancement is an important problem within the field of speech and signal processing, with impact on many computer-based speech recognition systems as well as coding and communication applications[1], [2]. Existing approaches to this task include traditional methods such as spectral subtraction[3], [4], Wiener filtering[4], [5], and Ephraim-Malah filtering[6]. Recently, wavelet shrinkage is a simple denoising technique based on the thresholding of the wavelet coefficients (WCs) and has emerged as a powerful tool for removing noise from signal in many signal-processing applications[7]–[10]. Donoho et al.[7], [8] proposed a universal threshold for removing the additive white Gaussian noise. By adequately choosing wavelet coefficient threshold (WCT), the corrupting white Gaussian noise can be efficiently removed by subtracting a threshold from noisy WCs. However, the method may not work well in enhancing colored-noise corrupted signal. Since then, adaptive wavelet-based methods in speech enhancement have been widely developed[9], [11]. They utilize variant WCT to improve the performance of speech enhancement. Bahoura et al.[11] proposed a method of threshold adaptation in time domain. Utilizing a Teager energy operator (TEO) to improve the discriminability of whether a speech frame is speech-dominated or noise-dominated. However, this method is only based on time adaptation of WCT. It utilizes the same threshold in each subband and suffers from serious residual noise and speech distortion in colored-noise infected speech. In addition, their method still existed over thresholding problem in speech enhancement applications[9].

Recently, some techniques have been proposed to reduce the effect of muscial residual noise[12], [13]. Virag[12] made use of masking properties of the human auditory system to reduce the effect of residual noise. Since human ears cannot perceive additive noise when at levels below the noise masking threshold (NMT). The methods that adopt the masking property of the human auditory system can reduce the effect of musical residual noise, but the drawback is the large computational effort associated with the subband decomposition and the additional FFT analyzer required for psychoacoustic modeling.

The objective of this paper is to introduce a novel speech enhancement based on time-frequency adaptation for providing robustness to non-stationary and colored noise. First, the bark-scale wavelet packet decomposition (BSWPD) is used to emulate the subband decomposition of human ear operations on auditory signals. In order to let WCT be temporally adjusted with the time varying noise levels for a non-stationary noise, a voice-activity detector (VAD) approach is then utilized to control the noise estimate update in speech enhancement system. In this paper, an accurate VAD approach is almost required if the background is not stationary or the SNR in the noisy signal is low. Traditionally, the VAD algorithms use short-term energy, zero-crossing rate and LPC coefficients[28] as feature parameters for detecting voice activity segment (VAS). Cepstral features[29], formant shape[30], and least-square periodicity measure[31] are some of the more recent metrics used in VAD designs. A statistical-model-based VAD[26] is also

Manuscript revised October 10, 2009.
†The author is with the Department of Information Technology & Communication, Shin Chien University, No. 200, University Rd, Neimen Shiang, Kaohsiung 845, Taiwan.
 a) E-mail: wkc0224@seed.net.tw
 DOI: 10.1587/transinf.E93.D.341

Copyright © 2010 The Institute of Electronics, Information and Communication Engineers
one of the VAD algorithms. The standardized VAD algorithms [19], [26] are recently proposed. On the other hand, the method based on higher-order statistics has been proposed to utilize the statistics of speech signals for VAD [32]. In order to let the VAD decision is unconcerned with variable of noise-level, the VAD is based on entropy measure. Then, we adopt the theory of adaptive frequency subband extraction (AFSE) to only use the few frequency subbands which are slightest corrupted by noise and discard the seriously obscured ones. A measure of entropy defined on the frequency subbands are selected by AFSE to form a robust feature for voiced speech. Moreover, unvoiced information is, moreover, eliminated from conventional wavelet-based VAD algorithm. In the most techniques which use the wavelet thresholding for speech enhancement, they may not only suppress additional noise but also some speech components like unvoiced ones. In order to overcome this drawback, a method of unvoiced decision is proposed in this paper. Finally, an inverse BSWPD is applied to re-synthesize the enhanced speech.

The rest of this paper is organized as follows. Section 2 describes the wavelet denoising techniques. Section 3 demonstrates the utilized entropy-based VAD approach. Section 4 derives the proposed speech enhancement algorithm. Section 5 describes the implementation of the proposed algorithm. The results are presented and discussed in Sect. 6. Section 7 concludes this paper.

2. Wavelet De-Noising Techniques

2.1 Wavelet Transform

Wavelet transform is based on time-frequency signal analysis [22]. It adopts a windowing technique with variable-sized regions. It allows the use of long time intervals when a more precise low-frequency information is desired and shorter regions where the high-frequency information is important. It is well known that speech signals contain many transient components and exhibit non-stationary properties. When we make use of the multi-resolution analysis (MRA) property of the Wavelet transform, high levels of time-resolution in the high frequency range is needed to detect signals’ rapid changing transient components. Conversely, a better frequency resolution is needed in the low frequency range to track slowly varying formants more precisely. Through MRA analysis, the classification of speech into voiced, unvoiced or transient components can be accomplished. Wavelet de-noising algorithm has recently emerged as a powerful tool for removing noise from signal in many signal processing applications.

2.2 Bark-Scale Wavelet Packet Decomposition

Critical subband is widely used in perceptual auditory modeling [14]. In this section, we propose the wavelet tree structure of bark-scale wavelet packet decomposition (BSWPD) to mimic the time-frequency analysis of the critical subbands according to the hearing characteristics of human cochlea. The bark-scale wavelet packet decomposition (BSWPD) is used to decompose the speech signal \( x(n) \) into 24 critical wavelet subband signals,

\[
wc^j_{\xi,m}(l) = BSWPD(x(n))
\]

where \( wc^j_{\xi,m}(l) \) defines the wavelet coefficients (WCs) with index \( l \) corresponding to the \( \xi \)th critical subband in the level \( j \) at \( m \)th frame. \( n \) is used as the discrete-time index for sampling points.

The tree structure of the proposed BSWPD can be constructed as shown in Fig. 1. It is implemented with an efficient five-level tree structure. Observing the Fig. 1, the whole band is decomposed into 24 critical subbands by using the high-pass filter and low-pass filter [15], where the symbol \( \downarrow 2 \) denotes an operator of downsampling by 2. The filter band is implemented with the Daubechies family wavelet as shown in Fig. 2.

2.3 Denoising by Wavelet Thresholding

Donoho and Johnstone proposed their original denoising method [8], which processed by thresholding wavelet coefficients of artificial signals. They attempt to recover a signal from noisy data with a Gaussian white noise, and they proposed a universal WCT \( \lambda \) for the wavelet transform:

\[
\lambda = \sigma \sqrt{2 \log(N_{frm})}
\]
where \( \sigma = \text{MAD}/0.6745 \) means the noise level. \( \text{MAD} \) represents the median of the absolute value of WCs. \( N_{\text{frm}} \) is used as the discrete-time index for sampling points.

Generally, the corrupting noise is not almost a white signal. The traditional wavelet-based speech enhancement methods which have been proposed in literatures assume that the WCT on each subband is the same.

The level-dependent threshold \( \lambda_j \) [17] was defined by

\[
\lambda_j = \sigma_j \sqrt{2 \log(N_{\text{frm}})}.
\]

(3)

where \( \sigma_j = \text{MAD}_j/0.6745 \). \( \text{MAD}_j \) represents the absolute median estimation of WCs on the level \( j \).

In the wavelet packet transform (WPT) case, the threshold is defined as

\[
\lambda = \sqrt{2 \log(N_{\text{frm}})}
\]

(4)

3. The Approach of VAD in Wavelet Domain

To design effective wavelet-based denoising algorithm for enhancing speech under non-stationary noise environment, it needs a well performance of VAD approach which estimates the noise spectrum accurately to determine different wavelet shrinkage during speech and non-speech segments for speech enhancement. In addition, voiced/unvoiced (V/UV) categories should be integrated into the VAD system to prevent degradation of unvoiced portions in the wavelet shrinking of noisy speech. In this section, a novel VAD comprises three parts: 1. AFSE; 2. BS-Entropy; 3. Unvoiced Decision and is shown in Fig. 3.

3.1 Calculation of Entropy Defined in Bark-Scale Frequency Bank

Originally, the entropy was defined for information sources by Shannon [23]. It measures the average length of bit code per symbol under optimal coding and is defined as:

\[
\text{Entropy}(S) = \sum_{k=1}^{K} P(s_k) \cdot \log_2\left(\frac{1}{P(s_k)}\right)
\]

(5)

where \( S = \{s_k\}_{0 \leq k \leq N-1} \) and \( P(s_k) \) is the probability of symbol \( s_k \). The entropy represents the amount of uncertainty in a source of \( K \) symbols. Based on this concept, the estimation of the entropy in spectral domain has been proposed for the application of speech detection [20]. It is found that the value of the entropy in the spectrum domain during speech segment is bigger than that during non-speech segment. The measure of entropy is defined in the spectral energy domain as:

\[
\text{Entropy}\left(\{X(\omega, t)\}^2\right) = \sum_{\omega=1}^{N} P(\{X(\omega, t)\}^2) \cdot \log_{10}\left(\frac{1}{P(\{X(\omega, t)\}^2)}\right)
\]

(6)

where \( P(\{X(\omega, t)\}^2) = \frac{|X(\omega, t)|^2}{\sum_{\omega=1}^{N} |X(\omega, t)|^2} \) is the probability of the frequency bin \( \omega \) for the magnitude spectrum for frame \( t \). The value is obtained by normalizing the summation of magnitude spectrum for all frequency bins.

Considering entropy defined in bark-scale wavelet domain, we calculate the probability \( P(\xi, m) \) of each bark-scale subband \( \xi \) for the \( m \)th frame as follow:

\[
P(\xi, m) = \frac{\text{WE}(\xi, m)}{\sum_{\omega=1}^{24} \text{WE}(\omega, m)}.
\]

(7)

where \( \text{WE}(\xi, m) = \sum_{l=1}^{N} |\text{WC}_{\xi,m}(l)|^2 \) means the wavelet energy of the \( \xi \)th subband for the \( m \)th frame.

Based on this probability \( P(\xi, m) \), the bark-scale spectral entropy, called as BS-Entropy parameter, is derive as follows,

\[
\text{BS-Entropy}(m) = \sum_{\xi=1}^{24} P(\xi, m) \cdot \log_{10}\left(\frac{1}{P(\xi, m)}\right).
\]

(8)

Some frequency subbands, however, are corrupted seriously by additive noise, and those harmful subbands may result in incorrect speech detection using BS-Entropy parameter if those harmful frequency subbands are considered. To solve this problem, we extract only the useful frequency subbands to calculate a measure of entropy defined on selected frequency subbands. The probability associated with subband
energy modified from Eq. (7) is described as follows:

\[ P(\xi, m) = \frac{WE(\xi, m)}{N_{\text{ab}}(m)} \sum_{\omega=1}^{N_{\text{ab}}(m)} WE(\omega, m) \]  

(9)

where \( N_{\text{ab}}(m) \) means the number of useful frequency subband which vary with frame index \( m \). Having finishing applying the above constraints, the BS-Entropy can be redefined below.

\[ \text{BS-Entropy}(m) = \frac{\sum_{\xi=1}^{N_{\text{ab}}(m)} P(\xi, m) \cdot \log_{10} \frac{1}{P(\xi, m)}}{ \sum_{\omega=1}^{N_{\text{ab}}(m)} WE(\omega, m) } \]  

(10)

The derivation of the BS-Entropy implies that the parameter depends only on the variation of the spectral energy but not on the amount of spectral energy. This BS-Entropy can calculate the uncertainty of spectral magnitude. Consequently, the BS-Entropy parameter is robust against changing level of noise, and we can use it to improve the noise estimation in the speech segment.

3.2 Adaptive Frequency Subband Extraction (AFSE)

The seriously obscured frequency subbands have little signal information, which is harmful for the results of VAD. Based on the finds, in our algorithm we must accurately extract only the useful frequency subbands to perform a VAD approach. How to select the useful subbands, however, is crucial for the VAD approach. Since our goal is to select some useful frequency subbands having the maximum signal information, we need a parameter to stand for the amount of useful signal information of each frequency subband. In general, the pure speech signal is an easy and good indicator.

The frequency subbands energy of pure speech signal is accomplished by removing the frequency energy of background noise from the frequency energy of input noisy speech.

The \( \xi \)th frequency subbands energy of pure speech signal of the \( m \)th frame \( \overline{WE}(\xi, m) \)

\[ \overline{WE}(\xi, m) = WE(\xi, m) - \sigma^2_n(\xi, m), \]  

(11)

where \( \sigma^2_n(\xi, m) \) is the estimated noise power of the \( \xi \)th frequency subband. To accurately update noise spectrum, the subband noise power, \( \sigma^2_n(\xi, m) \) can be recursively estimated by averaging past spectral power values using a time and frequency dependent smoothing parameter and be discussed later.

It is found that the more the frequency subband covered by noise would result in the smaller the \( \overline{WE}(\xi, m) \). Since the frequency subband with higher \( \overline{WE}(\xi, m) \) contains more pure speech information, we should sort the frequency subband according to their \( \overline{WE}(\xi, m) \) value.

That is,

\[ \overline{WE}(I_1, m) \geq \overline{WE}(I_2, m) \geq \cdots \geq \overline{WE}(I_N, m), \]

where \( I_n \) is the index of the frequency subband with the \( n \)th max energy.

It means that the index of the frequency subband with higher energy is the more useful index of one. Moreover, we should only select the useful frequency subbands for VAD results output. That is, the first \( N \) frequency subbands \( I_1, I_2, \ldots, I_N \) are selected and denoted as the useful number of frequency subband, \( N_{\text{ab}} \), for the succeeding calculation of spectral entropy. According to the relation between the number of useful frequency subbands \( N_{\text{ab}} \) and \( \text{SNR}_{\text{frm}} \) (shown as Fig. 4), we can see that the number of useful frequency subband varies with the value of \( \text{SNR}_{\text{frm}} \) under three types noises including white noise, factory noise and vehicle noise. \( N_{\text{ab}} = 9 \) and \( N_{\text{ab}} = 24 \) denote the boundary of \( N_{\text{ab}} \) among the range from -5 dB to 30 dB, respectively. Based on the above finds, a linear function can be used to simulate the relationship between \( N_{\text{ab}} \) and \( \text{SNR}_{\text{frm}} \), shown as Fig. 5.

\[
N_{\text{ab}}(m) = \begin{cases} 
9, & \text{SNR}_{\text{frm}}(m) < -5 \text{ dB} \\
(24 - 9) \times \frac{\text{SNR}_{\text{frm}}(m) - (-5)}{30 - (-5)} + 9, & \text{SNR}_{\text{frm}}(m) \leq 30 \text{ dB} \\
24, & \text{SNR}_{\text{frm}}(m) > 30 \text{ dB}.
\end{cases}
\]

(13)

Fig. 4 The results of correct detection accuracy with number of different frequency subband at -5 dB, 10 dB and 30 dB under three types of noise.

Fig. 5 A linear function of the relationship between \( N_{\text{ab}} \) and \( \text{SNR}_{\text{frm}} \).
where \([\cdot]\) is the round off operator. \(SNR_{frm}(m)\) denotes a frame-based posterior SNR for the \(m\)th frame.

\[
SNR_{frm}(m) = \sum_{n=1}^{24} SNR_{sub}(\xi, m),
\]

where \(SNR_{sub}(\xi, m)\) is decided by the useful subband SNR and is discussed later.

Observing Eq. (14), the value of \(SNR_{frm}(m)\) is dependent on the summation of the all useful critical subband \(SNR_{sub}(\xi, m)\).

3.3 Unvoiced Decision

Unlike voiced speech sounds, unvoiced speech sounds do not have any component of harmonic frequency. The majority of unvoiced sounds, however, display string spectral concentration in higher frequency range. The background noise display uniform spectral distribution. By a measure of energy distribution, unvoiced speech can be discriminate from background noise.

First, the average energy is calculated for each wavelet subband. By accumulating the average energy of the subbands below 2 kHz, we can compute the energy of lower frequency subband where the speech signal mainly focuses and show as below:

\[
E_{L,0}(m) = \sum_{j=1}^{8} wc_{j,m}^5, \quad E_{L,1}(m) = \sum_{j=9}^{12} wc_{j,m}^4,
\]

\[
E_{L,2}(m) = \sum_{j=13}^{18} wc_{j,m}^4 + wc_{19,m}^3,
\]

\[
EL(m) = \sum_{k=0}^{2} E_{L,k}(m),
\]

where \(EL(m)\) means the total energy from 0 Hz to 2 kHz. \(E_{L,0}(m)\) denotes the energy from 0 Hz to 0.5 kHz. \(E_{L,1}(m)\) denotes the energy from 0.5 kHz to 1 kHz, and \(E_{L,2}(m)\) denotes the energy from 1 kHz to 2 kHz.

If \(E_{L,2} > E_{L,1} > E_{L,0}\), it may be the existence of the voiced sound [27]. Similarly, energy of high-bands (EH) of the segment can be calculated by accumulating the average energy of the bands above 2 kHz.

\[
E_{H,0}(m) = \sum_{j=20}^{21} wc_{j,m}^3, \quad E_{H,1}(m) = \sum_{j=22}^{24} wc_{j,m}^2,
\]

\[
EH(m) = \sum_{k=0}^{1} E_{H,k}(m).
\]

The fact that the unvoiced sounds focus on higher frequency subband than voiced sounds, so the energy ratio of low frequency to high frequency is further used to determine the unvoiced segment:

\[
U(m) = \begin{cases} 
1, & \text{if } E_{L,2} > E_{L,1} > E_{L,0} \text{ and } EL/EH < 0.9 \\
0, & \text{otherwise.}
\end{cases}
\]

3.4 The Derivation of Voice Activity Segment

Finally, the decision of voice activity segment \(VAS(m)\) is performed with OR operating derived as:

\[
VAS(m) = BS-Entropy(m) \cup U(m),
\]

where \(\cup\) means the computation of OR logic.

4. Proposed Speech Enhancement Algorithm

The architecture of adaptive wavelet denoising system employing BS-entropy VAD approach is shown in Fig. 6.

A noisy speech signal, \(x(n)\), is modeled as:

\[
x(n) = s(n) + w(n),
\]

where \(s(n)\) and \(w(n)\) represent clean speech signal and background noise signal, respectively. The fact that the background noise level varies with time, the noise tracking plays a major role in determining the quality of a speech enhancement system, especially in non-stationary environment. The value of VAS parameter obtained from VAD approach is used to calculate the subband noise power. The noise estimation for each subband is computed using the adaptive noise estimation algorithm proposed by Lin et al. [24].

\[
\tilde{\sigma}_w^2(\xi, m) = \alpha(\xi, m) \cdot \tilde{\sigma}_w^2(\xi, m - 1) + [1 - \alpha(\xi, m)] \cdot WE(\xi, m)
\]

where

\[
\alpha(\xi, m) = \begin{cases} 
1 & \text{if } VAS(m) = 0, \\
\frac{1}{1 + e^{-k \cdot (SNR_{sub}(\xi, m) - T)}} & \text{if } VAS(m) = 1.
\end{cases}
\]

is smoothing parameter which is chosen as a sigmoid function, and its value varies with the estimate of a posterior signal-to-noise ratio during a non-speech segments (if \(VAS(m) = 0\)). Otherwise, the smoothing parameter will be set one during a speech segment (if \(VAS(m) = 1\)). \(k\) and \(T\) are the slope and center-offset of the sigmoid function respectively. Elevating \(k\) can decrease the transition range according to posteriori subband SNR. On the contrary, decreasing it would increase the transition range. During the
initialization period, the noisy signal is assumed to be noise-only and the noise spectrum is estimated by averaging the initial 10 frames.

The result of noise tracking can be used to calculate the subband-based SNR:

$$SNR_{sub}(\xi, m) = 10 \cdot \log_{10} \frac{WE(\xi, m)}{\sigma_w^2(\xi, m - 1)}$$

where $\sigma_w^2(\xi, m - 1)$ is the estimated noise power of the previous frame. The value of $SNR_{sub}(\xi, m)$ is determined by the ration of the observed $\xi$th subband wavelet energy to the previous $\xi$th subband estimated noise power. Consequently, the $SNR_{sub}(\xi, m)$ parameter will help us sense how much the current subband is corrupted by noise. Therefore, we will use this information for denoising.

In general, a frame with large number of high $SNR_{sub}$ implies that the current frame is a speech-dominated frame. On the contrary, a frame with a few number of small $SNR_{sub}$ implies that the frame is either in a noise-only region or in a very noisy environment. For a speech-dominated frame, the wavelet threshold of the frame should be made smaller. The wavelet coefficients are contributed mostly by the noise component in a noise-dominated frame. Thus, we propose a novel scheme that adjusts WCT according to the value of subband-based SNR, $SNR_{sub}(\xi, m)$, and formulate the WCT as below:

$$\lambda(\xi, m) = \lambda_j \left[ 1 - \frac{1}{1 + e^{-k(SNR_{sub}(\xi, m) - T)}} \right]$$

(23)

The speech-dominated frame can be further categorized into two types those are the voiced speech and the unvoiced speech according to the U/V decision. A voiced frame possesses a strong tone-like spectrum in lower subbands, so that the WCs of lower frequency must be reserved. On the contrary, the WCT tends to increase in lower frequency if the frame is categorized as unvoiced speech.

The voiced sounds are quasi-periodic in the time domain and harmonically structured. In frequency domain, these sounds are generally localized in bands that are less than 1 kHz. For many vowels of male and female voices, the statistic results indicate approximately that the frequency of the first formant doesn’t exceed 1 kHz and is superior to 0.1 kHz. Consequently, when a voiced-dominated frame form U/V decision, the WCT from Eq. (23) must be adapted to as

$$\lambda'(\xi, m) = \begin{cases} \alpha_L \cdot \lambda(\xi, m), & \text{if } \xi > 0.1 \text{ kHz and } \xi < 1 \text{ kHz} \\ \alpha_H \cdot \lambda(\xi, m), & \text{otherwise} \end{cases}$$

(24)

where $\alpha_j = 0.1$ and $\alpha_H = 1.0$ are experimentally determined. The frequency boundary covers most of the tone-like frequency components.

However, the energy of the unvoiced sounds is usually concentrated in high frequencies ($\geq 3$ kHz). If an unvoiced-dominated frame form U/V decision, the WCT from Eq. (23) must be adjusted to as

$$\lambda'(\xi, m) = \begin{cases} \beta_H \cdot \lambda(\xi, m), & \text{if } \xi > 3 \text{ kHz} \\ \beta_L \cdot \lambda(\xi, m), & \text{otherwise} \end{cases}$$

(25)

where $\beta_L = 1.2$ and $\beta_H = 0.05$ are experimentally determined.

The higher subbands contain less voiced information, reducing the WCs in higher subbands would suppress background noise. The higher subbands contain more significant information than the lower subbands do in an unvoiced frame. Hence, preserving the WCs of higher subbands can achieve a better performance by reducing the WCT in higher wavelet subbands shown as Eq. (25). The WCs corresponding to the lower subbands must be reduced to suppress the background noise.

The enhanced speech signal is synthesized with the inverse transformation of wavelet coefficients after thresholding:

$$\tilde{s}(n) = BSWPD^{-1}\{wc'_{\xi, m}\}$$

(26)

where $BSWPD^{-1}$ means inverse BSWPD.

5. Implementation of the Proposed Algorithm

Here we summarize the proposed algorithm by the following pseudo code:

Set frame size $L_{fm}$, and segment size $L_{seg}$ for $m = 1$ to $L_{seg}$

Split the whole band into 24 bark-scale frequency subband and obtain WCs on each subband by Eq. (1)

Calculate the wavelet energy $WE(\xi, m)$ on the $\xi$th sub-band

if $m \leq 10$

Assume as noise-only frame, and set $VAS(m) = 0$

Let noise power $\sigma_w^2(\xi, m)$ equal to wavelet energy $WE(\xi, m)$

Set $N_{sub}(m) = 9$

Compute the two thresholds of BS-Entropy($m$) and $EL/EH(m)$ using Eqs. (10), (15) and (16) respectively.

Set $\alpha(\xi, m) = 0$

Let time-adapted wavelet coefficient threshold equal to level-dependent wavelet coefficient threshold:

$\lambda(\xi, m) = \lambda_j$

if ($m == 10$)

$$\sigma_w^2(\xi, m) = \frac{1}{m} \sum_{k=0}^{9} \sigma_w^2(\xi, m - k)$$

end

else

Compute $SNR_{sub}(\xi, m)$ using Eq. (22)

Compute $SNR_{frm}(m)$ using Eq. (14), and determine $N_{sub}(m)$ according to Eq. (13)

Obtain the BS-Entropy($m$) using Eq. (10)
Estimate the ratio of $EL/EH(m)$ using Eqs. (15)–(16). Determine the voice activity segment (VAS) using Eq. (18). Compute $\alpha(\xi, m)$ dominated by sigmoid function using Eq. (21). Estimate the current noise power $\tilde{\sigma}_w^2(\xi, m)$ by averaging past spectral power value which uses a time and frequency dependent smoothing parameter $\theta(\xi, m)$ using Eq. (20). Adjust the WCT by using Eq. (23). The WCT can be adjusted again using Eq. (24) when a voiced-dominated frame is discovered. The WCT also be adjusted again using Eq. (25) if an unvoiced-dominated frame is found.

6. Experimental Results

Due to the VAD approach is important in the proposed wavelet-based speech enhancement, a number of experiments are evaluated. To set up the noisy signal for test, we add the prepared noise signals to the recorded speech signal with different SNRs range from $-5$ dB to 10 dB. Noise signals extracted from the Noisex-92 database [18] were artificially added to clean speech signals with various SNRs. Of the various noises available on the NOISEX database, white noise, factory noise and vehicle noise are selected as speech contamination.

6.1 The Evaluation of VAD

In order to compare with other VADs, we introduce three criteria: 1) the probability of correctly detecting speech frames $P_{cS}$ is the ratio of the correct speech decision to the total number of hand-labeled speech frames. 2) the probability of correctly detecting noise frames $P_{cN}$ is the ratio of the correct noise decision to the total number of hand-labeled noise frames. 3) the probability of false error $P_f$ is the ratio of the false speech decision or false noise decision to the total hand-labeled frames. The experimental results are summarized in Tables 1–3. Under a variety of SNR’s, the three criteria of proposed algorithm are compared with those of the VADs specified in the ITU standard G.729B [19], other entropy-based VAD proposed by Shen et al. [20] and HOS-based VAD [32]. It is shown that the proposed VAD algorithm obtains a better performance than those of others. The proposed VAD has superior performance to others particularly in low SNR, even though the result of Shen’s VAD is comparable to proposed VAD in high SNR. Figure 7 shows the VAD result of the proposed algorithm on the noisy speech signal “May-I-Help-you” under variable-level of noise. It is found that the VAS of the proposed algorithm can correctly extract speech segments especially for unvoiced segment /H/ occurred at /Help/ sentence in Fig. 7 (b). Conversely, in Fig. 7 (c) the

| Type | SNR(dB) | Proposed | G.729B [19] | Shen [20] | Noise (32) |
|------|---------|----------|-------------|-----------|------------|
| White | 10      | 99.4     | 93.1        | 98.1      | 92.5       |
|      | -5      | 93.6     | 88.2        | 94.6      | 83.6       |
| factory | 10     | 94.6     | 93.9        | 94.5      | 93.7       |
|      | -5      | 89.7     | 84.5        | 95.1      | 84.2       |
| vehicle | 10     | 90.6     | 75.6        | 94.8      | 76.8       |
|      | -5      | 78.6     | 67.4        | 81.2      | 69.8       |

6.2 The Evaluation of Speech Enhancement

The performance of the proposed speech enhancement system based on time-frequency adapted WCTs is evaluated in various adverse conditions and compared to those of the others [9], [11]. It has been shown by experiments that even though the SNRs are very similar at the output of the enhancement system, the listening test and speech spectrograms can produce very divergent results. In our experiments, the subjective evaluation and objective evaluation are applied to validate the performance of the speech enhancement method. We select the speech database contained 60 speech phrases (in Mandarin and in English) spoken by 32 native speakers (22 males and 10 females), sampled at 16 kHz with 16-bit resolution. The following experimental results, speech enhancement was conducted with the experimentally chosen parameter values: $L_{frm} = 256$, $N = 24$, $\alpha_L = 0.1$, $\alpha_H = 1.0$, $\beta_L = 1.2$, $\beta_H = 0.05$, $k = 0.2$ and $T = 1.5$.

6.2.1 Subjective Evaluation

A five-scale absolute opinion from 1 (poor) to 5 (excellent)
was adopted to subjectively evaluate the 20 listeners. The scale used for these tests corresponds to the mean opinion score (MOS) scale [21]. The score can represent the global perception of the residual noise, background noise and speech distortion. The subjective listening tests are performed by subjecting the speech to varying noise at SNR = 0 and 15 dB and the results are presented in Table 4. In the proposed system, we apply an appropriate threshold into the unvoiced sound for improving the intelligibility of speech signal. The subjective listening tests show that the proposed enhancement method produces the highest quality speech perceived by the actual human listeners among the algorithms being tested especially for low SNR.

### 6.2.2 Objective Evaluation

Table 5 presents the average SegSNR results of the speech enhancement evaluations for different methods. The average SegSNR improvements are used for the performance evaluations in different noise environments. The higher average SegSNR results show that the proposed algorithm has much better enhancement performance than others. This is the reason that the WCT of each subband is then temporally adjusted for improving the speech quality in the presence of noise according to the result of BS-Entropy-based VAD. To further evaluate the proposed speech enhancement algorithm for non-stationary noise, the experimental results (objective evaluation) in Table 6. Observing Table 6, the 60 speech phrases were mixed with the variable-level of factory and vehicle noises. Even though the values of SegSNR almost decrease in Table 6, the results show that the values of SegSNR of proposed method are also higher than other two methods for variable-level of factory and vehicle noises. Owing to the utility of BS-Entropy-based VAD, the WCT can be adjusted properly during speech-dominated frame especially in variable-level of background noise.

### 7. Conclusion

In the paper, the proposed speech enhancement is developed using time-frequency wavelet threshold. The proposed method employs entropy defined in bark-scale wavelet packet decomposition parameter to estimate the noise level. The experiments reveal that a VAD approach us-
ing BS-Entropy parameter has more excellent presentation than other VADs especially in variable-level background noise. By employing BS-Entropy parameter, our proposed method outperforms those other wavelet-based methods and has been shown to be able to efficiently remove background noise. Simulation results show that the proposed system is capable of reducing noise with little speech degradation especially in variable noise-level, and the overall performance is superior to several competitive methods in both objective and subjective evaluations.

Acknowledgements

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC 98-2221-E-158-004.

References

[1] B.H. Juang, “Recent developments in speech recognition under adverse conditions,” Proc. Int. Conf. Spoken Language Process ‘90, pp.1113–1116, 1990.
[2] J.H. Chen and A. Gersho, “Adaptive postfiltering for quality enhancement of coded speech,” IEEE Trans. Speech Audio Process., vol.3, no.1, pp.57–71, 1995.
[3] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-24, no.2, pp.113–120, 1979.
[4] J.R. Deller, J.H.L. Hansen, and J.G. Proakis, Discrete-Time Processing of Speech Signals, second ed., IEEE Press, New York, 2000.
[5] S. Haykin, Adaptive Filter Theory, third ed., Prentice Hall, Upper Saddle River, New Jersey, 1996.
[6] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-32, no.6, pp.1109-1121, 1984.
[7] D.L. Donoho, “De-noising by soft thresholding,” IEEE Trans. Inf. Theory, vol.41, no.3, pp.613–627, May 1995.
[8] D.L. Donoho and I.M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage,” Biometrika, vol.81, no.3, pp.425–455, 1994.
[9] S.H. Chen and J.F. Wang, “Speech enhancement using perceptual wavelet packet decomposition and teager energy operator,” J. VLSI Signal Processing, vol.36, pp.125–139, 2004.
[10] S.F. Lei and Y.K. Tung, “Speech enhancement for nonstationary noises by wavelet packet transform and adaptive noise estimation,” Proc. 2005 International Symposium on Intelligent Signal Processing and Communication Systems, pp.41–44, Dec. 2005.

Kun-Ching Wang was born in Kaohsiung, Taiwan in 1976. He received the B.S. degree in electrical engineering from Southern Taiwan University in 1998 and M.S. degree in electrical engineering from Feng Chia University in 2000, and the Ph.D. degree in control engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2005. From 2005 to Jan. 2008, he was R&D in ITRI. Since Feb. 2008, he has been with the Department of Information Technology & Communication, Shin Chien University, Kaohsiung, Taiwan, where he is currently an Assistant Professor. His research interests include adaptive signal processing, speech recognition, speech enhancement, wavelet analysis and applications.