Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 924135, 8 pages
doi:10.1155/2009/924135

Research Article

Wavelet-Based Speech Enhancement Using Time-Frequency Adaptation

Kun-Ching Wang

Department of Information Technology & Communication, Shin Chien University, No. 200, University Road, Neimen Shiang, Kaohsiung 845, Taiwan

Correspondence should be addressed to Kun-Ching Wang, wkc0224@seed.net.tw

Received 22 February 2009; Revised 21 July 2009; Accepted 11 October 2009

Recommended by Satya Dharanipragada

Wavelet denoising is commonly used for speech enhancement because of the simplicity of its implementation. However, the conventional methods generate the presence of musical residual noise while thresholding the background noise. The unvoiced components of speech are often eliminated from this method. In this paper, a novel algorithm of wavelet coefficient threshold (WCT) based on time-frequency adaptation is proposed. In addition, an unvoiced speech enhancement algorithm is also integrated into the system to improve the intelligibility of speech. The wavelet coefficient threshold (WCT) of each subband is first temporally adjusted according to the value of a posterior signal-to-noise ratio (SNR). To prevent the degradation of unvoiced sounds during noise, the algorithm utilizes a simple speech/noise detector (SND) and further divides speech signal into unvoiced and voiced sounds. Then, we apply appropriate wavelet thresholding according to voiced/unvoiced (V/U) decision. Based on the masking properties of human auditory system, a perceptual gain factor is adopted into wavelet thresholding for suppressing musical residual noise. Simulation results show that the proposed method is capable of reducing noise with little speech degradation and the overall performance is superior to several competitive methods.

1. Introduction

Many speech signal processing applications have been applied in real-world [1]. The performance of speech coding and recognition system that operate in noisy environments decrease when high ambient noise levels occur. Therefore, speech enhancement system becomes a hot research topic to improve the performance of many computer-based speech recognition systems, coding and communication applications [2, 3]. The exiting methods such as spectral subtraction [4, 5], Wiener filtering [5, 6], and Ephraim-Malah filtering [7] are well-known. Recently, wavelet shrinkage has emerged as a powerful tool for removing noise from signal [8–11]. It is a simple denoising technique based on the thresholding of the wavelet coefficients (WCs). Donoho and Johnstone firstly proposed a universal threshold for removing the additive white Gaussian noise [8, 9]. In addition, they also proposed a level-dependent threshold to remove colored noise [12]. Bahoura and Rouat proposed a method of threshold adaptation in time domain by utilizing the use of Teager energy operator (TEO) [13]. The TEO can improve the discriminability for a speech frame. Chen et al. presented an improved wavelet-based speech enhancement method using the perceptual wavelet packet decomposition and the TEO. Lu and Wang proposed a method that the background noise can be almost removed by adjusting the wavelet coefficient threshold (WCT) according to the value of SNR [14]. After that, the adaptive wavelet-based methods in speech enhancement are widely presented. They utilize adequately WCT to improve the performance of speech enhancement.

For noisy speech, energies of unvoiced segments are comparable to those of noise. 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. Consequently, the detection of the voiced/unvoiced segments of the speech signals is a main problem in wavelet-based methods. Sheikhzadeh and Abutalebi [15] suggested an improved scheme, which categorized speech into either a voiced frame or an unvoiced frame. They increased WCT for high bands in a voiced frame and decreased the threshold values for high bands in an unvoiced frame. As a result, both low-frequency components

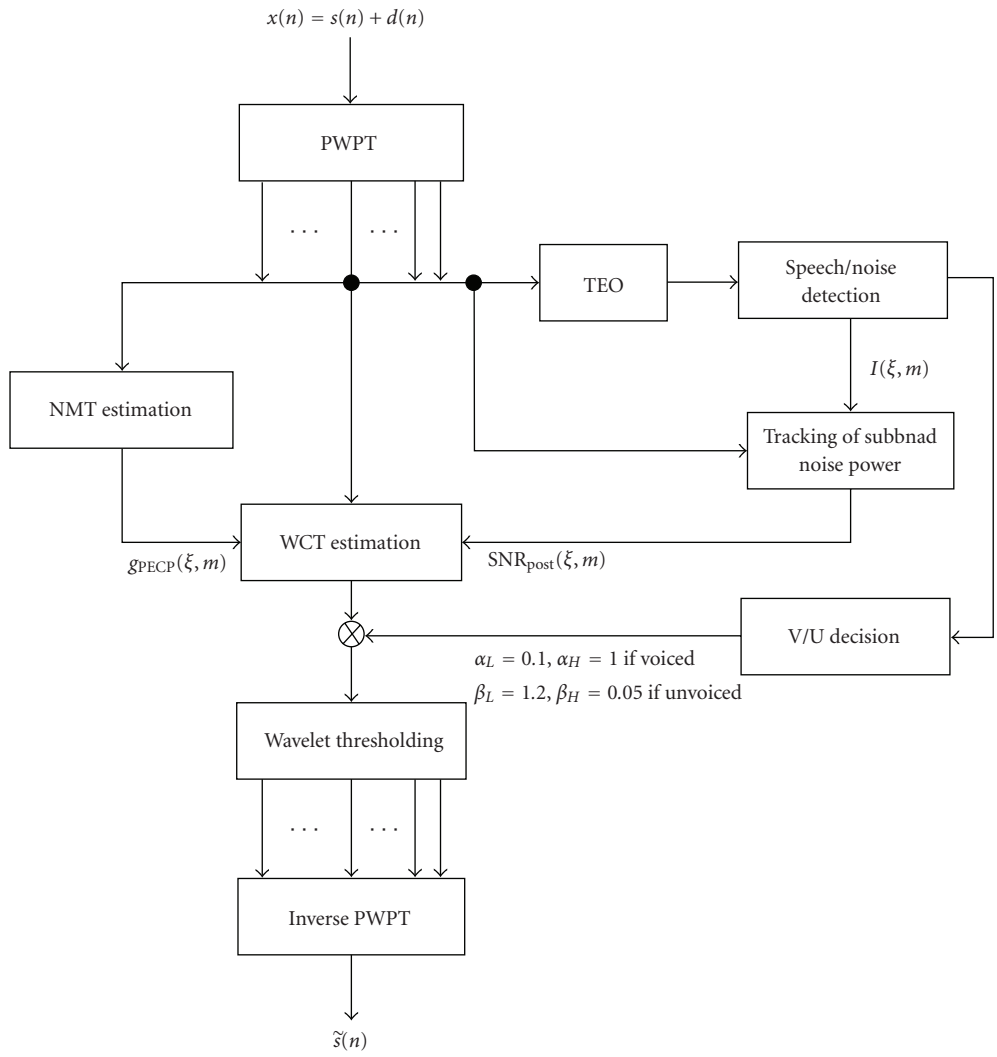

Figure 1: The architecture of proposed speech enhancement method based on the time-frequency adaptation of the wavelet threshold.

of voiced segments and high-frequency components of unvoiced segments are reserved by the soft thresholding algorithm. In addition, a number of methods were considered to reduce the effect of musical residual noise [16, 17]. Since human ears cannot perceive additive noise when at levels below the noise masking threshold (NMT), Virag used the masking properties of the human auditory system to suppress the effect of musical residual noise [16].

In this paper, we introduce a novel wavelet-based speech enhancement using time-frequency adaptation for providing robustness to nonstationary and colored noise. The perceptual wavelet packet transform (PWPT) is applied to approximate the human auditory system. The wavelet coefficient threshold (WCT) of each subband is first temporally adjusted according to the value of a posterior signal-to-noise ratio (SNR). Consequently, utilizing V/U decision, the different threshold values are used as voiced and unvoiced frames to further improve the intelligibility of the processed speech signal. In addition, the musical residual noise can be efficiently suppressed to improve the perceptual quality when a gain factor is typically derived according to the NMT. Finally, an inverse PWPT is applied to resynthesize the enhanced speech.

2. Proposed Speech Enhancement Algorithm

Let $s(n)$ represent a discrete time speech signal, and let $d(n)$ denote a discrete time background noise signal. The noise-corrupted speech signal $x(n)$ can be modeled as $x(n) = s(n) + d(n)$. The architecture of proposed speech enhancement method based on the time-frequency adaptation of the wavelet threshold is shown in Figure 1, and the proposed method is organized in the following seven steps.

2.1. Perceptual Wavelet Packet Transform (PWPT). Critical subband is widely used in perceptual auditory modeling [18]. In this work, a perceptual wavelet packet transform (PWPT) is used to decompose the speech signal from 20 Hz to 16 kHz into 24 critical frequency subbands:

$$w_{\xi}^{j}(k) = \text{PWPT}\{x(n)\}, \quad \xi = 1, \ldots, 24, \tag{1}$$

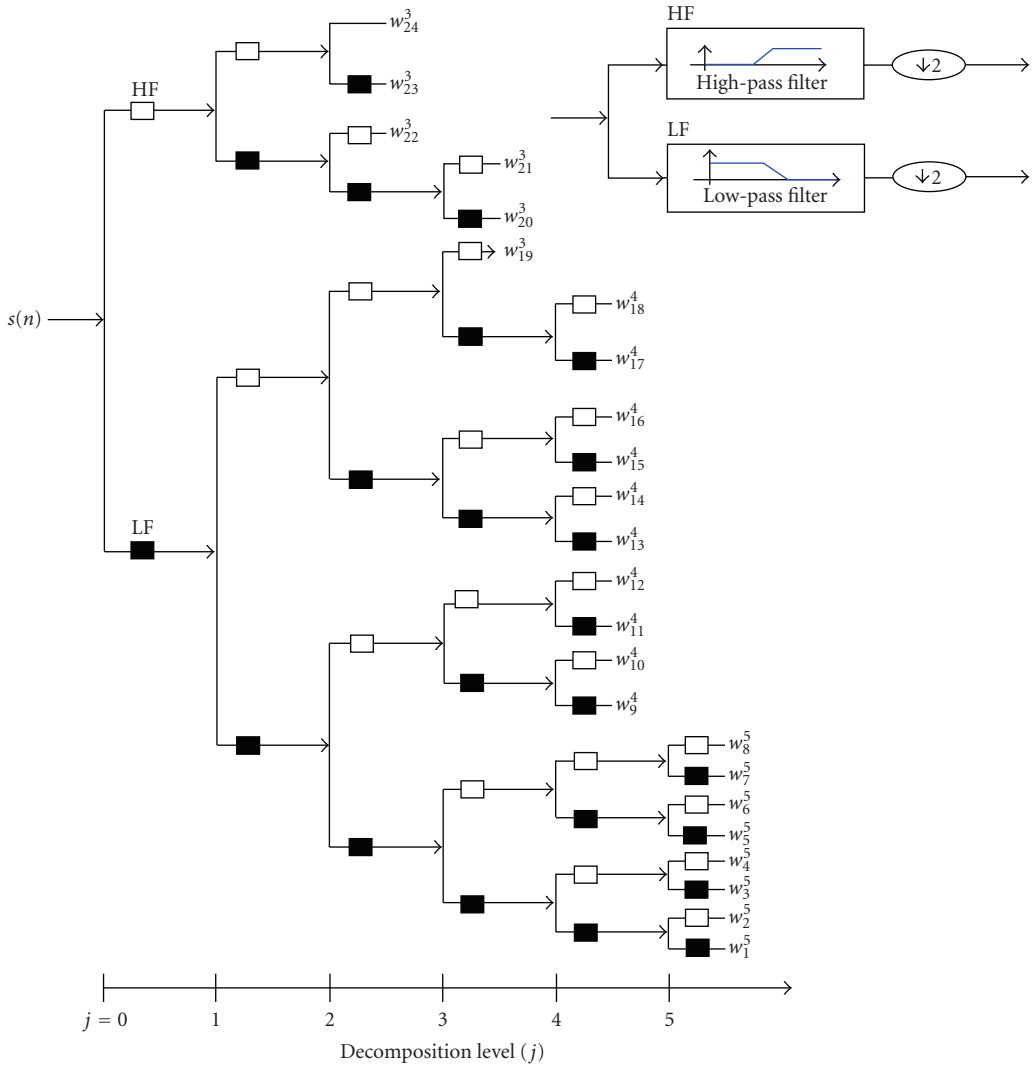

Figure 2: The structure of wavelet packet transform (PWPT).

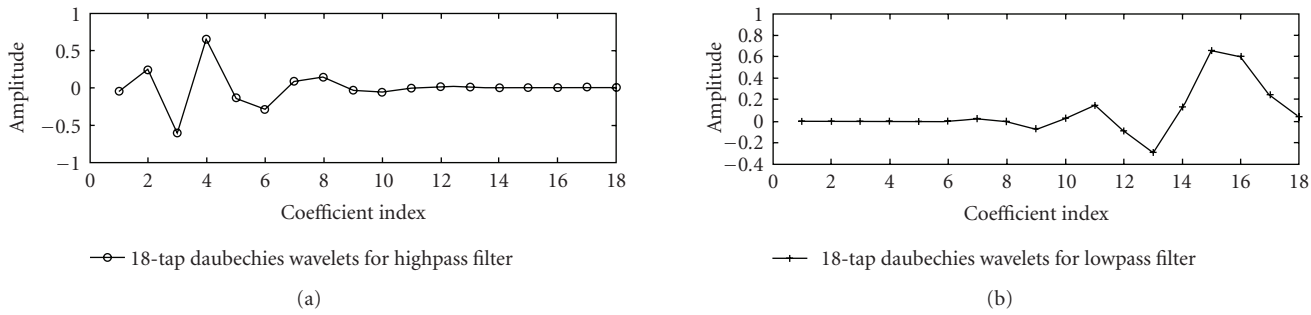

Figure 3: The highpass filter and lowpass filter implemented with the Daubechies family wavelet.

where $w_\xi^j(k)$ means the kth coefficient of the ξth subband on level j. PWPT$\{\cdot\}$ denotes a process of PWPT.

Figure 2 shows the implementation of an efficient five-level tree structure. Before an operator of downsampling by 2 in each level, the lowpass (LP) and highpass (HP) are implemented with 18-tap FIR filters derived from the Daubechies family wavelet shown in Figure 3 [19].

2.2. Speech/Noise Detector (SND) Using Teager Energy on Wavelet Domain. Various techniques for detecting voiced/unvoiced (V/U) speech regions have been proposed; however, the performance of the speech/noise detector (SND) is dramatically degraded in noise. The teager energy operator (TEO) is a powerful nonlinear operator; it has been experimentally observed that the TEO can enhance

the discriminability among speech and noise and further suppress the noise components from noisy speech signals [20]. The discrete-time TEO is applied to the wavelet coefficients $w^j_{\xi,m}(k)$:

$$t^j_{\xi,m}(k) = w^j_{\xi,m}(k)^2 - w^j_{\xi,m}(k-1) \cdot w^j_{\xi,m}(k+1), \quad (2)$$

where m represents the frame index.

The simple SND algorithm computes the level 1 energy on wavelet coefficients of discrete-time TEO, $t^j_{\xi,m}(k)$. If the percentage of energy concentrated in level 1 approximation is above 90% of the total energy, the current frame is regarded as speech-dominated segment:

$$\mathrm{SND}(m) = \begin{cases} \text{voiced}, & \dfrac{\left[\sum_k t^1_{1,m}(k)\right]^2}{\left[\sum_k t^1_{1,m}(k) + \sum_k t^1_{2,m}(k)\right]^2} > 0.9, \\ \text{unvoiced/noise}, & \text{otherwise}, \end{cases} \quad (3)$$

where $t^1_{1,m}(k)$ and $t^1_{2,m}(k)$ are the Teager coefficients of approximation and detail subband, respectively.

To further separate the unvoiced sound from noise segments, a method of unvoiced decision is proposed in this section. According to the tree structure of PWPT (shown in Figure 2), the three subenergies corresponding to the wavelet subband signals are defined as

$$\begin{aligned} E_{L0}(m) &= \sum_{\xi=1}^{8} \sum_k t^5_{\xi,m}(k), \\ E_{L1}(m) &= \sum_{\xi=9}^{12} \sum_k t^4_{\xi,m}(k), \\ E_{L2}(m) &= \sum_{\xi=13}^{18} \sum_k t^4_{\xi,m}(k) + \sum_k t^3_{19,m}. \end{aligned} \quad (4)$$

The unvoiced frame on mth frame is determined as

$$\mathrm{SND}(m) = \begin{cases} \text{unvoiced}, & \text{if } E_{L2}(m) > E_{L1}(m) > E_{L0}(m), \\ & \dfrac{E_{L0}(m)}{E_{L2}(m)} < 0.99, \\ \text{noise}, & \text{otherwise}. \end{cases} \quad (5)$$

2.3. The Tracking of Subband Noise Power. Since the background noise level varies with time, the tracking of noise plays a major role in determining the quality of a speech enhancement system, especially in nonstationary environment. The decision result from SND approach is used to update the subband noise power. Then, the subband noise power, $\widetilde{\sigma}^2_d(\xi, m)$, can be adaptively estimated by [21]

$$\widetilde{\sigma}^2_w(\xi, m) = \widetilde{\alpha}_d(\xi, m) \cdot \widetilde{\sigma}^2_d(\xi, m-1) + [1 - \widetilde{\alpha}_d(\xi, m)] \cdot \varepsilon(\xi, m), \quad (6)$$

where $\widetilde{\alpha}_d(\xi, m) = \alpha_d + (1 - \alpha_d) \cdot p'(\xi, m)$, $p'(\xi, m) = \alpha_p \cdot p'(\xi, m-1) + (1 - \alpha_p) \cdot I(\xi, m)$. $\varepsilon(\xi, m)$ is the energy of ξth critical subband and is defined in later. $\widetilde{\alpha}_d(\xi, m)$, $\alpha_d(\xi, m)$, and $\alpha_p(\xi, m)$ all represent the smoothing parameter. $p'(\xi, m)$ and $I(\xi, m)$ are a conditional signal presence probability and an indicator of voice-dominated, respectively.

Observing (6), $I(\xi, m)$ is an indicator of updating noise power. The parameter depends on the speech-present ratio and is determined by the decision of speech-only frame. If $\mathrm{SND}(m)$ = voiced or unvoiced sounds, let $I(\xi, m) = 1$. Consequently, $\widetilde{\alpha}_d(\xi, m)$ is increasing and the noise power of next frame is nearly updated from the current estimated noise power. Conversely, $\mathrm{SND}(m)$ = noise period, let $I(\xi, m) = 0$. Consequently, $\widetilde{\alpha}_d(\xi, m)$ is decreasing and the noise power of next frame is nearly updated from the current observed signal power.

The result of noise tracking can be used to calculate a posterior signal-to-noise ratio (SNR):

$$\mathrm{SNR}_{\mathrm{post}}(\xi, m) = 10 \cdot \log_{10} \frac{\varepsilon(\xi, m)}{\widetilde{\sigma}^2_d(\xi, m-1)}, \quad (7)$$

where $\widetilde{\sigma}^2_d(\xi, m-1)$ is the estimated noise power of the previous frame. The value of $\mathrm{SNR}_{\mathrm{post}}(\xi, m)$ is determined by the ratio of the observed ξth subband wavelet energy to the previous ξth subband estimated noise power. Consequently, the $\mathrm{SNR}_{\mathrm{post}}(\xi, m)$ parameter will help us sense how much the current subband is corrupted by noise. Therefore, we will use this information for denoising noise. During the initialization period, the observed power is assumed to be noise only and the noise spectrum is estimated by averaging the initial 10 frames.

2.4. Estimation of Noise Masking Threshold (NMT). This subsection describes the incorporation of the human auditory masking properties into our enhancement system. The NMT is estimated on the WCs of PWPT. At first, WCs are obtained from the PWPT of noisy speech. The energy of ξth critical subband is calculated by

$$\varepsilon(\xi, m) = \sum_{l(\xi)}^{h(\xi)} \left| w^j_{\xi,m}(k) \right|^2, \quad (8)$$

where $l(\xi)$ and $h(\xi)$ are the coefficient indices of the first and last wavelet coefficients in ξth critical subband [16].

An excitation pattern $B(\xi, m)$ can be regarded as an energy distribution along the basilar membrane. $B(\xi, m)$ can be calculated by convolving the subband energy $\varepsilon(\xi, m)$ with the spreading function $F(\xi)$ given by [16, 22]

$$B(\xi, m) = F(\xi) * \varepsilon(\xi, m) \quad (9)$$

A relative threshold offset $O(\xi)$, which can be found in [12, 16], specifies whether a speech frame is tone like or noise like. This threshold should be imposed when adjusting the log subband energy. Therefore, a threshold $\widetilde{B}(\xi, m)$ is computed as the sum of the log energy for the excitation

pattern and the offset $O(\xi)$, written as

$$\widetilde{B}(\xi, m) = 10 \cdot \log_{10} B(\xi, m) + O(\xi), \tag{10}$$

where the values of the offset $O(\xi)$ are all negative.

Convolving the subband energy $\varepsilon(\xi, m)$ with the spreading function $F(\xi)$ increases the energy in each subband, so to multiply each $\widetilde{B}(\xi, m)$ by the inverse of the energy gain is necessary for renormalization. Accordingly, a normalized threshold is given by

$$\mathrm{Th}(\xi, m) = \widetilde{B}(\xi, m) - G(\xi, m), \tag{11}$$

where $G(\xi, m)$ denotes the gain factor between the spread energy $B(\xi, m)$ and the subband energy $\varepsilon(\xi, m)$ in dB.

$G(\xi, m)$ is expressed as

$$G(\xi, m) = 10 \cdot \log_{10}\left(\frac{B(\xi, m)}{\varepsilon(\xi, m)}\right). \tag{12}$$

Additionally, the normalized threshold $\mathrm{Th}(\xi, m)$ is compared with the absolute-hearing threshold (AHT) which is frequency-dependent and can be closely approximated as [16, 22]

$$\mathrm{AHT}(f) = 3.64 f^{-0.8} - 6.5 e^{-0.6(f-3.3)^2} + 0.001 f^4 \,[\mathrm{dB}] \tag{13}$$

with f in kHz.

Finally, the NMT $T(\xi, m)$ is obtained by

$$T(\xi, m) = \max\{\mathrm{AHT}(f), \mathrm{Th}(\xi, m)\}, \tag{14}$$

where f is chosen as the central frequency of the critical band ξ.

2.5. Estimation of Wavelet Coefficient Threshold. In this work, we propose a novel scheme that adjusts WCT according to the value of a posterior SNR and formulate the WCT as follows:

$$\lambda(\xi, m) = \lambda_j \cdot \left(1 - \frac{1}{1 + e^{-\gamma \cdot (\mathrm{SNR}_{\mathrm{post}}(\xi, m) - \eta)}}\right), \tag{15}$$

where $\lambda_j = \mathrm{MAD}_j/0.6745 \cdot \sqrt{2 \cdot \log(N_j)}$ means the level-dependent threshold λ_j [12]. MAD_j represents the absolute median estimated at the jth level. γ and η are the slope and center-offset of the Sigmoid function, respectively. These two factors are chosen to be 0.2 and 1, respectively.

Observing (15), the value of $\lambda(\xi, m)$ is adjusted by a Sigmoid functions, and its value varies with the estimate of a posterior signal-to-noise ratio while locating nonspeech segments. Otherwise, the smoothing parameter will be set one. γ and η are the slope and center-offset of the Sigmoid function, respectively. Elevating γ can decrease the transition range according to posteriori subband SNR. On the contrary, decreasing it would increase the transition range.

In general, a frame with high value of signal-to-noise ratio (SNR) implies that the current frame is a speech-dominated frame. On the contrary, a frame with low value of SNR implies that the frame is either in a noise-only region or in a very noisy environment. So, the wavelet threshold of the frame should be made smaller for a speech-dominated frame. The wavelet coefficients are contributed mostly by the noise component in a noise-dominated frame.

The speech-dominated frame can be further categorized into two types: those are the voiced speech and the unvoiced speech according the V/U 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 quasiperiodic 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 does not exceed 1 kHz and is superior to 100 Hz. Consequently, when a voiced-dominated frame forms V/U decision, the WCT from (15) must be adapted to as

$$\lambda'(\xi, m) = \begin{cases} \alpha_L \cdot \lambda(\xi, m), & \text{if } f_\xi > 100\,\mathrm{Hz},\ f_\xi < 1000\,\mathrm{Hz}, \\ \alpha_H \cdot \lambda(\xi, m), & \text{otherwise}, \end{cases} \tag{16}$$

where $\alpha_L = 0.1$ and $\alpha_H = 1.0$ are experimentally determined. The frequency boundary covers most of the tone-like frequency components. f_ξ denotes the frequency bin of subband ξ.

In (16), more WCs in lower subbands must be properly reserved since a voiced frame contains strong tone-like components in the lower frequency. This can be accomplished by reducing the WCT in lower wavelet subbands.

However, the energy of the unvoiced sounds is usually concentrated in high frequencies ($\geq$3 kHz). If an unvoiced-dominated frame forms V/U decision, the WCT from (15) must be adjusted to as

$$\lambda'(\xi, m) = \begin{cases} \beta_H \cdot \lambda(\xi, m), & \text{if } f_\xi > 3000\,\mathrm{Hz}, \\ \beta_L \cdot \lambda(\xi, m), & \text{otherwise}, \end{cases} \tag{17}$$

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, reserving the WCs of higher subbands can achieve a better performance by reducing the WCT in higher wavelet subbands shown as (17). The WCs corresponding to the lower subbands must be reduced to suppress the background noise.

In order to improve the final perceptual quality after thrsholding, a suppression method of musical residual noise can adopt a perceptual gain factor into wavelet thresholding. The time-frequency-adapted wavelet threshold is modified from (16) and (17), respectively:

$$\lambda''(\xi, m) = \lambda'(\xi, m) \cdot g_{\mathrm{PECP}}(\xi, m), \tag{18}$$

Table 1: The objective evaluation using SegSNR improvement.

Noise type	Method	Input SNR [dB] −5	0	5	10
White Gaussian	TI	5.13	4.11	2.6	2.37
	WPD+TEO	9.52	6.74	4.42	2.48
	TSA	7.84	6.28	4.02	2.61
	PER+WPT	10.52	8.44	5.63	3.58
	Proposed	**13.52**	**11.06**	**9.02**	**5.62**
Vehicle	TI	7.43	5.02	2.79	2.04
	WPD+TEO	8.76	5.84	4.52	2.51
	TSA	7.11	4.82	3.03	2.02
	PER+WPT	9.34	6.73	5.64	3.52
	Proposed	**12.01**	**9.84**	**8.62**	**5.42**
Factory	TI	3.48	3.23	1.68	1.41
	WPD+TEO	5.43	3.54	2.52	1.82
	TSA	4.53	2.84	2.03	1.32
	PER+WPT	5.84	4.52	3.06	2.01
	Proposed	**9.22**	**8.51**	**6.02**	**4.92**

Table 2: The objective evaluation using IS measure.

Noise type	Method	Input SNR [dB] −5	0	5	10
White Gaussian	TI	3.48	3.1	2.68	2.52
	WPD+TEO	2.2	2.01	1.74	1.59
	TSA	2.32	2.19	1.99	1.76
	PER+WPT	2.21	2.02	1.77	1.6
	Proposed	**1.94**	**1.62**	**1.51**	**1.3**
Vehicle	TI	3.62	3.39	3.13	2.55
	WPD+TEO	1.71	1.65	1.48	1.29
	TSA	2.11	1.91	1.73	1.68
	PER+WPT	1.98	1.71	1.73	1.42
	Proposed	**1.4**	**1.22**	**1.07**	**0.92**
Factory	TI	3.71	3.49	3.26	2.71
	WPD+TEO	1.96	2.81	1.63	1.42
	TSA	2.23	2.12	1.94	1.7
	PER+WPT	2.11	2.02	1.72	1.65
	Proposed	**1.63**	**1.31**	**1.22**	**1.02**

where $g_{\mathrm{PECP}}(\xi,m) = 1/(1+\max\{\sqrt{|\widetilde{\sigma}_d^2(\xi,m)|/T(\xi,m)}-1,0\})$ denotes a perceptual gain factor given by [23], and $T(\xi,m)$ is derived from the NMT.

From (18), it is known that if the energy of musical residual noise, $\widetilde{\sigma}_d^2(\xi,m)$, is greater than the NMT in a subband, the wavelet coefficient thresholds become small adjusted by the gain factor to suppress infecting noise. However, if the energy of residual noise is smaller than the NMT, the corrupting noise cannot be perceived by the human ear. We do not need to change the WCTs for retaining the speech quality.

2.6. Soft Thresholding. The noise components are suppressed by soft thresholding wavelet packet coefficients of the noisy signal as follows:

$$\widetilde{w}_{\xi,m}^{j}(k) = \begin{cases} \mathrm{sgn}\left[w_{\xi,m}^{j}(k)\right] \cdot \left[\left|w_{\xi,m}^{j}(k)\right| - \lambda''(\xi,m)\right], \\ \qquad \text{if } \left|w_{\xi,m}^{j}(k)\right| > \lambda''(\xi,m), \\ 0, \qquad \text{if } \left|w_{\xi,m}^{j}(k)\right| \le \lambda''(\xi,m), \end{cases} \tag{19}$$

Table 3: The objective evaluation using PESQ measure.

Noise type	Method	Input SNR [dB] −5	0	5	10
White Gaussian	TI	1.10 (1.04)	1.37 (1.24)	1.88 (1.55)	2.10 (1.96)
	WPD+TEO	1.25 (1.04)	1.45 (1.24)	1.94 (1.55)	2.26 (1.96)
	TSA	1.26 (1.04)	1.42 (1.24)	1.92 (1.55)	2.18 (1.96)
	PER+WPT	1.28 (1.04)	1.72 (1.24)	2.01 (1.55)	2.30 (1.96)
	Proposed	**1.62 (1.04)**	**1.85 (1.24)**	**2.13 (1.55)**	**2.45 (1.96)**
Vehicle	TI	1.21 (1.04)	1.42 (1.24)	1.89 (1.55)	2.12 (1.96)
	WPD+TEO	1.39 (1.04)	1.52 (1.24)	1.92 (1.55)	2.31 (1.96)
	TSA	1.33 (1.04)	1.48 (1.24)	1.90 (1.55)	2.10 (1.96)
	PER+WPT	1.48 (1.04)	1.51 (1.51)	1.84 (1.55)	2.41 (2.16)
	Proposed	**1.83 (1.04)**	**1.95 (1.51)**	**2.46 (1.55)**	**2.50 (2.16)**
Factory	TI	1.19 (1.04)	1.33 (1.24)	1.92 (1.55)	2.01 (1.96)
	WPD+TEO	1.28 (1.04)	1.41 (1.24)	1.98 (1.55)	2.15 (1.96)
	TSA	1.21 (1.04)	1.38 (1.24)	1.92 (1.55)	2.10 (1.96)
	PER+WPT	1.47 (1.04)	1.51 (1.42)	2.28 (1.55)	2.34 (2.11)
	Proposed	**1.59 (1.04)**	**1.78 (1.42)**	**2.31 (1.55)**	**2.49 (2.11)**

where $\mathrm{sgn}[\cdot]$ is the sign function. $\tilde{w}^{j}_{\xi,m}(k)$ is thresholded wavelet coefficient.

2.7. Inverse PWPT. Finally, the speech signal is synthesized with the inverse transformation of the thresholded wavelet packet coefficients as follows:

$$\tilde{s}(n) = \mathrm{PWPT}^{-1}\left\{\tilde{w}^{j}_{\xi,m}(k)\right\}, \tag{20}$$

where $\mathrm{PWPT}^{-1}\{\cdot\}$ means process of inverse PWPT.

3. Experimental Results

In this section, we select the speech database that contains 60 speech phrases (in Chinese Mandarin and in English) spoken by both male and female speakers. To set up the noisy signal for test, we added the prepared noise signals to the recorded speech signal with different SNRs range from −5 dB to 10 dB. A variety of nonstationary noises are taken from the Noisex-92 database [24] for experiments. All noisy signals are sampled at 8 kHz with 16 bits/sample. The frame size is 64 milliseconds and the frame shift is 32 milliseconds. To evaluate the performance of our algorithm, the methods including (1) speech enhancement using time-invariant threshold (TI) [8], (2) speech enhancement using perceptual wavelet packet decomposition and Teager energy operator proposed (WPD+TEO) [10], (3) wavelet speech based on time-scale adaptation (TSA) [25], and (4) speech enhancement method using perceptually constrained gain factors in critical-band-WPT proposed (PER+WPT) [14] are compared to our proposed algorithm.

Several objective speech quality measures including segmental SNR (SegSNR) improvements, Itakura-Saito (IS) measure [26], and perceptual evaluation of speech quality (PESQ) [27–29] are tested to vary noise at the range $[-5, 10]$ dB in this section. Table 1 shows the SegSNR improvements of the speech enhancement evaluations for different methods. The amounts of noise reduction, residual noise, and speech distortion can be measured by SegSNR improvement. Observing Table 1, the SegSNR improvements are used for the performance evaluations in different noise environments. The higher SegSNR improvements results show that the proposed method has much better enhancement performance than others. In addition, the perceptual gain factor offers the best performance at the lower SNR inputs in the proposed method. Table 2 shows the Itakura-Saito (IS) measure results of the speech enhancement. The majority of IS results show that the proposed method has the lower spectral distortion values than those of other methods at different SNR levels for various nonstationary noises such as factory and vehicle. The results of PESQ scores are performed by the actual human listeners among the algorithms and presented in Table 3. In Table 3, the comments in brackets are the scores of PESQ without thresholding process. It is observed that the proposed enhancement method produces the better improvement of quality speech than other methods especially for low SNR.

4. Conclusion

The proposed speech enhancement algorithm uses time-frequency wavelet threshold instead of traditional invariant and time-variant threshold. The wavelet coefficient threshold is adjusted according to the value of a posterior SNR. In addition, the V/U decision lets the WCT be different with voiced frame or unvoiced frame. A residual musical noise is successfully suppressed when a perceptual gain factor is adopted into the estimation of WCT. Experimental results

show that the proposed method yields a higher improvement in SegSNR, lower IS measure, and higher PESQ scores than other methods under all tested environmental conditions.

Acknowledgment

This research was partially sponsored by the National Science Council, Taiwan, under contract no. NSC 98-2221-E-158-004.

References

[1] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[2] B. H. Juang, "Recent developments in speech recognition under adverse conditions," in *Proceedings of the International Conference on Spoken Language Process (ICSLP '90)*, pp. 1113–1116, 1990.

[3] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 59–71, 1995.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[5] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, NY, USA, 2nd edition, 2000.

[6] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 1996.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[8] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[9] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[10] S.-H. Chen and J.-F. Wang, "Speech enhancement using perceptual wavelet packet decomposition and Teager energy operator," *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 36, no. 2-3, pp. 125–139, 2004.

[11] S.-F. Lei and Y.-K. Tung, "Speech enhancement for nonstationary noises by wavelet packet transform and adaptive noise estimation," in *Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS '05)*, pp. 41–44, Hong Kong, December 2005.

[12] I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *Journal of the Royal Statistical Society. Series B*, vol. 59, no. 2, pp. 319–351, 1997.

[13] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the Teager energy operator," *IEEE Signal Processing Letters*, vol. 8, no. 1, pp. 10–12, 2001.

[14] C.-T. Lu and H.-C. Wang, "Enhancement of single channel speech based on masking property and wavelet transform," *Speech Communication*, vol. 41, no. 2-3, pp. 409–427, 2003.

[15] H. Sheikhzadeh and H. R. Abutalebi, "An improved wavelet based speech enhancement system," in *Proceedings of the European Conference on Speech Communication and Technology*, pp. 1855–1858, 2001.

[16] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[17] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 59–67, 2004.

[18] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer, New York, NY, USA, 1990.

[19] S. Mallat, "Multifrequency channel decomposition of images and wavelet model," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 2091–2110, 1989.

[20] F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," *IEEE Signal Processing Letters*, vol. 6, no. 10, pp. 259–261, 1999.

[21] L. Lin, W. H. Holmes, and E. Ambikairajah, "Adaptive noise estimation algorithm for speech enhancement," *Electronics Letters*, vol. 39, no. 9, pp. 754–755, 2003.

[22] B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms," *IEEE Transactions on Signal Processing*, vol. 47, no. 6, pp. 1622–1635, 1999.

[23] C.-T. Lu and H.-C. Wang, "Speech enhancement using perceptually-constrained gain factors in critical-band-wavelet-packet transform," *Electronics Letters*, vol. 40, no. 6, pp. 394–396, 2004.

[24] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[25] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on time-scale adaptation," *Speech Communication*, vol. 48, no. 12, pp. 1620–1637, 2006.

[26] S. Quackenbush, T. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[27] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ): the new ITU standard for end-to-end speech quality assessment—part I: time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[28] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," ITU-T Recommendation, 2003.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 2, pp. 749–752, Salt Lake, Utah, USA, May 2001.