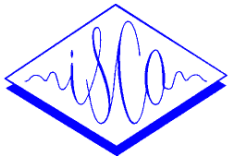

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

SPEECH ENHANCEMENT USING WAVELET TRANSFORM WITH CONSTRAINED THRESHOLDS

Ching-Ta LU

Department of Electrical Engineering,
National Tsing Hua University, Hsinchu
Department of Electronic Engineering,
Chin Min College, Miaoli
lucas1@ms26.hinet.net

Hsiao-Chung WANG

Department of Electrical Engineering,
National Tsing Hua University, Hsinchu
hcwang@ee.nthu.edu.tw

ABSTRACT

Although the power spectral subtraction (PSS) in frequency domain and the soft/hard thresholding function in wavelet domain have been successfully applied in speech enhancement, however they increase the probability of negative error and generate musical residual noise. In this paper, we propose a novel scheme that adapts the weight to wavelet coefficients (WCs) for each subband. The weights of all subbands form a weighting function, it can track the speech variation in each subband. Multiplying the weighting function to WCs, the enhanced speech is obtained by inverse wavelet transforming (IWT) the modified WCs. The experiments show that the proposed speech enhancement method can efficiently eliminate the additive noise. Moreover, it is almost free from the musical residual noise.

1. INTRODUCTION

Decomposing a signal in wavelet basis has been applied in image signal processing [1]. Due to the advantage of multi-resolution analysis, it can be utilized to analyze non-stationary random process, such as a speech signal [1~2]. Transforming a signal into wavelet domain, most of the wavelet coefficients (WCs) tend to be zero [3~4], this implies that wavelet representation can represent a signal with sparse nonzero coefficients. Thus it can be successfully applied in low-bit-rate speech coding [2].

In mobile communication, a user may send voice in noisy environment. To remove the background noise is an important task for mobile communication. Due to wavelet transform decomposes a signal in some subbands, it is adequate to analyze a rapid variation signal. Recently, many wavelet-based speech enhancement methods have been proposed to enhance noise-corrupted speech [2][4~8].

Donoho and Johnston [6] proposed the universal threshold to enhance noisy speech by the soft thresholding function. This method could efficiently suppress the background noise. It performed well in low SNR environments, however it is failed to enhance noisy speech in high SNR case. Bahoura and Rouat [8] suggested the level-dependent wavelet thresholding. This method can greatly reduce the noise, and also the time space adapted threshold can prevent the speech from quality deterioration during the thresholding process. The benefit of the method is that the SegSNR estimation is not necessary. Furthermore, they suggested that the WCT should be adequately adjusted according to articulation manner, i.e., voiced or unvoiced [9]. Jansen and Bultheel [10] suggested that the universal threshold should be multiplied by an adjustment factor to obtain the minimum risk. Kim, Yang and Kwon [4] suggested adaptive wavelet shrinkage for noise cancellation. They used entropy to part the noisy signal into unvoiced signal section and the other signal section, to apply different threshold value for each section. They applied soft thresholding function in unvoiced section, hard thresholding function in both voiced and speech-absence section. The experiments showed their method could efficiently denoising. Rezayee and Gazor [11] suggested an adaptive Karhunen-Loeve transform (KLT) tracking-based algorithm for enhancing colored-noise corrupted speech. Their algorithm decomposed noisy speech into its components along the axes of a KLT-based vector space of clean speech. Enhancing noisy speech by modifying each KLT component due to its noise and clean speech energies, the enhanced speech was obtained by inverse KLT.

In this paper, we derive a novel scheme which employs weighting function on wavelet coefficients to enhance noise-corrupted speech. When background noise has been efficiently suppressed, the musical residual noise can be almost immunized. Initially, the speech pause region is detected. If a frame is detected to be speech-absence, the WCs are saved as the estimated noise WCs. Since the variation of noise power is slow, the estimated noise WCs are not change until next speech-absence has been detected. Between two successive speech-absence regions, the variance of WCs of each subband is assumed to be unchanged. The estimated variance of WCs of clean speech can be obtained by subtracting estimated variance of noise WCs from the variance of noisy WCs. The weighting function is defined as the ratio of the variance of estimated speech WCs to the variance of noisy speech WCs for each subband. Then the enhanced speech WCs can be obtained by multiplying the weighting function to the noisy speech WCs.

Because the proposed weighting function is adapted for each subband, it works well in enhancing the colored-noise corrupted speech also. The experiments show that the enhanced speech of the proposed method sounds more natural than the subtraction-based method.

The rest of the paper is organized as follows. Section 2 describes the block diagram of the proposed speech enhancement system. Section 3 derives the weighting function of wavelet coefficients for each subband. The experimental results are demonstrated in Section 4. Finally, the conclusion is drawn in Section 5.

2. SPEECH ENHANCEMENT SYSTEM

The proposed speech enhancement system is depicted in Fig. 1. At first, the speech-pause detection is performed by energy-based method. If a speech-pause frame is detected, the variance of WCs of background noise is updated to track the noise variation. The weighting function corresponding to all of the subbands is estimated. The weighting function is multiplied to noisy speech WCs in each subband. Finally, the modified WCs are inversely transformed back to reconstruct the enhanced speech.

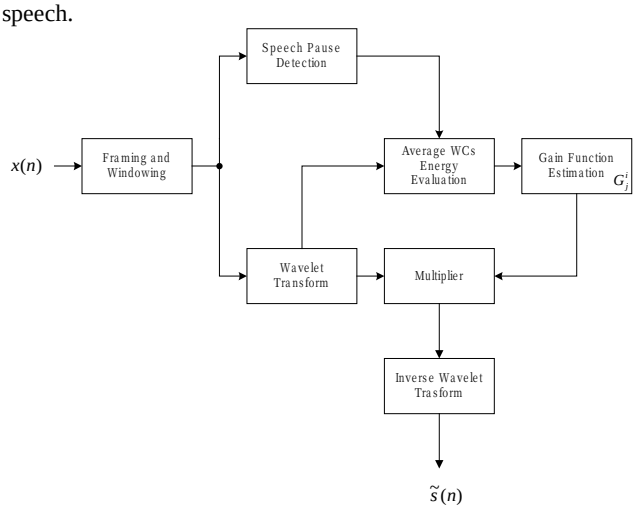

Fig. 1. The block diagram of proposed speech enhancement system.

3. DERIVATION OF THE PROPOSED WEIGHTING FUNCTION

A noisy speech signal can be modeled as the sum of clean speech and additive noise

$$x_m(n) = s_m(n) + w_m(n) \tag{1}$$

where $x_m(n)$, $s_m(n)$ and $w_m(n)$ denote the noisy speech, the clean speech, and the corrupting noise in *m*-th frame, respectively.

Taking the wavelet transform of the noisy speech $x_m(n)$, the wavelet coefficients of noisy speech can be expressed as the sum of wavelet coefficients of clean speech and those of noise.

$$X_{j,k}^i(m) = S_{j,k}^i(m) + W_{j,k}^i(m) \tag{2}$$

where $X_{j,k}^i$, $S_{j,k}^i$, and $W_{j,k}^i$ are the WCs of noisy speech, clean speech, and noise at *i*-th subband in 2^{-j} scale, respectively.

We proposed that the enhanced speech WCs are the multiplication of a weighting function to WCs for each subband. It can be written as

$$\tilde{S}_{j,k}^i(m) = G_j^i(m) \cdot X_{j,k}^i(m) \tag{3}$$

The wavelet coefficient distortion (WCD) is defined as the difference between clean speech WCs and enhanced speech WCs.

$$E_{j,k}^i(m) = \tilde{S}_{j,k}^i(m) - S_{j,k}^i(m) \tag{4}$$

Substitute (3) into (4), the WCD can be rewritten as

$$E_{j,k}^i(m) = G_j^i(m) \cdot X_{j,k}^i(m) - S_{j,k}^i(m) \tag{5}$$

Substitute (2) into (5), we can obtain

$$E_{j,k}^i(m) = S_{j,k}^i(m) \cdot [G_j^i(m) - 1] + G_j^i(m) \cdot W_{j,k}^i(m) \tag{6}$$

The WCD can be viewed as the sum of the WCD of clean speech $E_{S_{j,k}}^i$ and that of residual noise $E_{W_{j,k}}^i$.

$$E_{j,k}^i(m) = E_{S_{j,k}}^i(m) + E_{W_{j,k}}^i(m) \tag{7}$$

Since the contents of $s_m(n)$, $g_m(n)$ and $w_m(n)$ are all real, the conjugate operator can be omitted. The variance of WCD can be given as

$$E\{|\mathrm{E}_{j,k}^i(m)|^2\} = E\{|\mathrm{E}_{S_{j,k}}^i(m)|^2\} + E\{|\mathrm{E}_{W_{j,k}}^i(m)|^2\} + 2 \cdot E\{\mathrm{E}_{S_{j,k}}^i(m) \cdot \mathrm{E}_{W_{j,k}}^i(m)\} \tag{8}$$

Substitute (4) into (8), the variance of WCD can be rewritten as

$$\begin{aligned} E\{|\mathrm{E}_{j,k}^i(m)|^2\} = {} & E\{S_{j,k}^{i\,2}(m) \cdot [G_j^i(m) - 1]^2\} + E\{G_j^{i2}(m) \cdot W_{j,k}^{i\,2}(m)\} \\ & + 2 \cdot E\{\mathrm{S}_{j,k}^i(m) \cdot [G_j^i(m) - 1] \cdot G_j^i(m) \cdot W_{j,k}^i(m)\} \end{aligned} \tag{9}$$

In order to find the optimum weight of a subband, we differentiate (9) with respect to subband weight G_j^i , and let the result be zero. We can obtain

$$\begin{aligned} & 2 \cdot E\{S_{j,k}^{i\,2}(m) \cdot [G_j^i(m) - 1]\} + 2 \cdot G_j^i(m) \cdot E\{W_{j,k}^i(m)\}^2 \\ & + 2 \cdot E\{S_{j,k}^i(m) \cdot W_{j,k}^i(m)\} \cdot [2 \cdot G_j^i(m) - 1] = 0 \end{aligned} \tag{10}$$

Since $s_m(n)$ and $w_m(n)$ are mutually uncorrelated, the corresponding wavelet coefficients sequence are mutually uncorrelated also. WCs can be obtained by inner product of a signal and wavelet basis $\varphi_{j,k}^i(n)$.

$$S_{j,k}^i(m) = \langle s_m(n), \varphi_{j,k}^i(n) \rangle \tag{11}$$

$$W_{j,k}^i(m) = \langle w_m(n), \varphi_{j,k}^i(n) \rangle \tag{12}$$

The correlation between the WCs of clean speech and those of noise can be derived as

$$\begin{aligned} E\{S_{j,k}^i(m) \cdot W_{j,l}^i(m)\} &= E\left\{\sum_{n=0}^{N-1} s_m(n) \cdot \varphi_{j,k}^i(n) \cdot \sum_{u=0}^{N-1} w_m(u) \cdot \varphi_{j,l}^i(u)\right\} \\ &= \sum_{n=0}^{N-1} \sum_{u=0}^{N-1} E\{s_m(n)\} \cdot E\{w_m(u)\} \cdot \varphi_{j,k}^i(n) \cdot \varphi_{j,l}^i(u) \end{aligned} \tag{13}$$

Since $w_m(n)$ is an additive white Gaussian noise with zero mean, the correlation between the WCs of clean speech and those of noise equals to zero. That means (13) is equal to zero.

$$E\{S_{j,k}^i(m) \cdot W_{j,l}^i(m)\} = 0 \tag{14}$$

Substitute (14) into (10), the weighting function G_j^i is expressed by

$$G_j^i(m) = \frac{\mathrm{E}\{|S_{j,k}^i(m)|^2\}}{\mathrm{E}\{|S_{j,k}^i(m)|^2\} + \mathrm{E}\{|\mathrm{W}_{j,k}^i(m)|^2\}} \tag{15}$$

Since

$$\mu_{S_j}^{\ i}(m) = \mathrm{E}\left\{ S_{j,k}^{i}(m) \right\} = 0 \quad \text{and} \tag{16}$$

$$\mu_{W_j}^{\ i}(m) = \mathrm{E}\left\{ W_{j,k}^{i}(m) \right\} = 0 \ , \tag{17}$$

the variance of clean speech WCs of can be simplified as

$$\sigma^2_{S^i_{j,k}}(m) = \mathrm{E}\left\{ |S_{j,k}^{i}(m)|^2 \right\} \tag{18}$$

The variance of noise WCs can also be given as

$$\sigma^2_{W^i_{j,k}}(m) = \mathrm{E}\left\{ |W_{j,k}^{i}(m)|^2 \right\} \tag{19}$$

Substitute (18) and (19) into (15), the weighting function can be written as

$$G_j^i(m) = \frac{\sigma^2_{S_j^i}(m)}{\sigma^2_{S_j^i}(m) + \sigma^2_{W_j^i}(m)} \tag{20}$$

Obviously, the denominator of (20) equals to the variance of noisy speech WCs. (20) can be rewritten as

$$G_j^i(m) = \frac{\sigma^2_{\tilde{S}_j}(m)}{\sigma^2_{X_j^i}(m)} \tag{21}$$

Substituting (21) into (3) and inverse transforming back to the time domain, the enhanced speech is expressed by

$$\tilde{s}_m(n) = \sum_{j=L+1}^{J} \sum_{i=1}^{I_j} \sum_{k=0}^{2^{-j}-1} \frac{\sigma^2_{\tilde{S}_j}(m)}{\sigma^2_{X_j^i}(m)} \cdot X_{j,k}^{\ i}(m) \cdot \varphi_{j,k}^{i}(n) + \sum_{k=0}^{2^{-J}-1} \frac{\sigma^2_{\tilde{S}_J}(m)}{\sigma^2_{X_J}(m)} \cdot X_{J,k}(m) \cdot \phi_{J,k}(n) \tag{22}$$

where $\phi_{J,k}(n)$ denotes a scaling function at 2^{-J} scale.

In real environments, the variance of clean speech WCs are unknown. It can be estimated by subtracting the variance of background noise WCs from noisy speech WCs. The procedure is given as follows

$$E\left\{ |\tilde{S}_{j,k}^{i}(m)|^2 \right\} = \max\left\{ d_{j,k}^{i}(m) - E\{|\tilde{W}_{j,k}^{i}(m)|^2\}, 0 \right\} \tag{23}$$

4. EXPERIMNETS

In the experiments, the noisy speech is obtained by corrupting the clean speech with additive white Gaussian noise in three different noise levels. They are with segmental SNR equal to 0dB, 5dB and 10 dB. The test data are spoken in Mandarin with sampling frequency of 8 kHz, including six male and ten female speakers.

The noisy speech is not assumed to be stationary, but the variation of signal power is practically small in the successive frames. Thus, we interpolate the successive subbands the variance of WCs. That is to constrain the variation of the WCs. Then we can obtain the variance of WCs at 2^{-j} scale and *i-th* subband by

$$d_j^i(m) = \beta \cdot d_j^i(m-1) + (1-\beta) \cdot E\left\{ |X_{j,k}^{\ i}|^2 \right\} \tag{24}$$

The initial variance of noisy speech WCs $d_j^i(0)$ is set to be zero.

The weighting factor β is chosen to be 0.3.

In order to improve the performance of removing background noise in a speech-pause region, we adjust the weighting function by SegSNR of a frame. The modified weighting function can be given as

$$G_{j_{SNR}}^{i}(m) = \left[1 - 1/\left(1 + e^{-\alpha \cdot \hat{SNR}(m) + \gamma} \right) \right] \cdot G_j^i(m) \tag{25}$$

where α and γ are chosen to be 0.2 and 2, respectively.

We compare our proposed method with other methods, including the power spectral subtraction (PSS) and universal wavelet coefficient thresholding (UWCT) method [6]. In order to improve the performance of UWCT, we integrate the speech-absence detection algorithm with UWCT. The method is named the modified UWCT (MUWCT). The MUWCT method can still well work in the noisy environment of segmental SNR great than 10 dB. The comparison of SegSNR improvement is shown in Table 1. Obviously, the performance of our proposed method is better than that of others. It reveals that the proposed system benefits both low speech distortion and efficient noise reduction. Note that the above three speech enhancement methods use the same algorithm of speech-absence detection.

Table. 1. *Comparison of the improved SegSNR of enhanced speech for various noise levels.*

Noisy SegSNR (dB)	SegSNR of Enhanced Speech (dB)		
	PSS	MUWCT	Proposed
0	5.30	6.14	8.24
5	4.08	4.79	6.40
10	2.35	2.98	4.37

Fig.2. illustrates the spectrograms of clean speech, noisy speech, and the enhanced speech obtained from different enhancement methods. When the background noise has been efficiently eliminated by the proposed method, the enhanced speech can be free from musical residual noise. The waveforms of the enhanced speech are demonstrated in Fig. 3. It shows that our proposed method can efficiently remove the background noise in the speech-absence region. An informal listening test has been performed, the enhanced speech of the proposed scheme sounds more nature than the other two methods, and also the musical residual noise has been suppressed. The PSS method may produce musical residual noise. In lower SNR case, the MUWCT produces musical residual noise also.

5 CONCLUTIONS

The propose speech enhancement system has efficiently eliminated the background noise, and it can be almost free from the musical residual noise. We estimate a weighting factor for each subband by finding the ratio of estimated variance of clean speech WCs to the variance of noisy speech WCs. Inverse wavelet transforming the modified WCs back to time domain, the enhanced speech is obtained. This method works well in enhancing colored-noise infected speech also.

6. ACKNOWLEDGEMENT

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC-89-2614-E- 007-004.

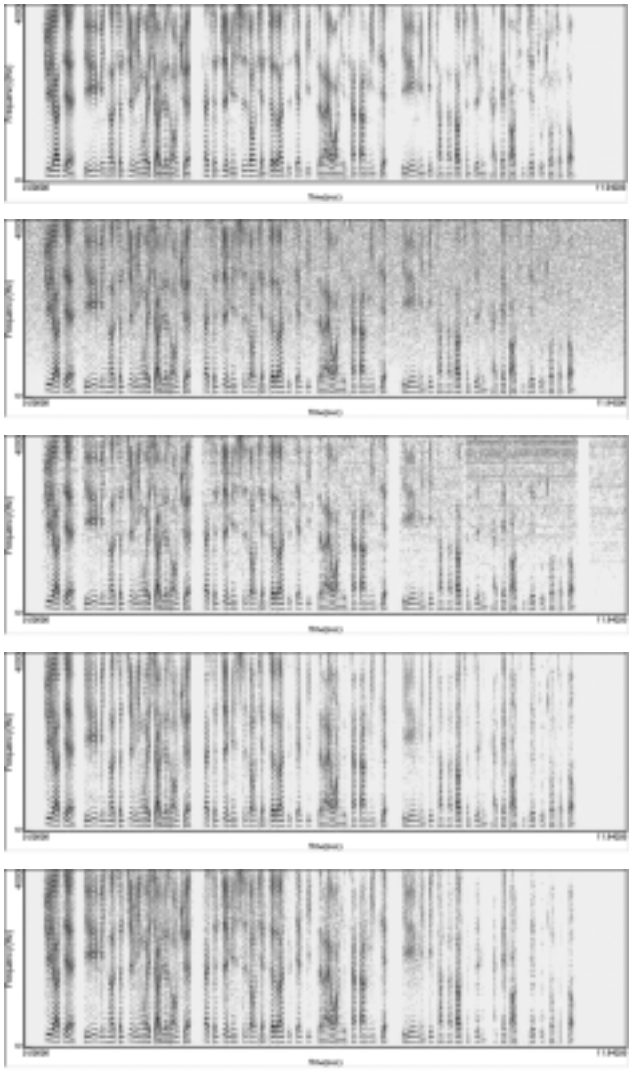

Fig. 2. *Spectrograms of clean speech (from top to bottom), noisy speech (SegSNR = 10 dB), enhanced speech using PSS, enhanced speech using MUWCT, and enhanced speech using the proposed method.*

7. REFERENCES

[1] Mallat S., A Wavelet Tour of Signal Processing, Academic Press, San Diego: A Harcourt Science and Technology, 1999.

[2] Carnero B. and A. Drygajlo "Perceptual Speech Coding and Enhancement Using Frame-Synchronized Fast Wavelet Packet Transform Algorithms," *IEEE Trans. Speech and Audio Processing*, VOL. 47, NO 6, p.p. 1622-1635, June 1999.

[3] Castrillon-Candas J. E. and Amaratunga K. "Fast Estimation of Continuous Karhunen-Loeve Eigenfunctions Using Wavelets," *IEEE Trans. Speech and Audio Processing*, VOL. 50, NO 1, p.p. 78-86, January 2002.

[4] Kim I. J., Yang S. I. and Kwon Y. "Speech Enhancement Using Adaptive Wavelet Shrinkage," *Proceedings of International Conference on Industrial Electronics*, Pusan, Korea, p.p. 501-504, 2001.

[5] Lu C. T. and Wang H. C. "Enhancement of Single Channel Speech Using Perception-Based Wavelet Transform," to be appeared in *Proceeding of International Conference on Spoken Language Processing*. Colorado, September 2002.

[6] Donoho D. L. and Johnston I. M., "Ideal spatial adaptation by wavelet shrinkage," Biometrika, vol. 81, no. 3, pp. 425-455, 1994.

[7] Singh L. and Sridharan S., "Speech enhancement for forensic application using dynamic time warping and wavelet packet analysis," in Proc. IEEE TENCON-SITCT, pp. 475-478, 1997.

[8] Bahoura M. and Rouat J., "Wavelet speech enhancement based on the teager energy operator," IEEE Signal Processing Letters, vol. 8, NO 1, pp. 10-12. Jan. 2001.

[9] Bahoura M. and Rouat J., "A new approach for wavelet speech enhancement, " in. Proc. European Conference on Speech Communication and Technology, pp. 1937-1940, 2001.

[10] Jansen M. and Bultheel A., "Asymptotic behavior of the minimum mean squared error threshold for noisy wavelet coefficients of piecewise smooth signals," IEEE Trans. Signal Processing, vol. 49, pp. 1113-1118, June 2001.

[11] Rezayee A. and Gazor S., "An Adaptive KLT Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, VOL. 9, NO 2, February 2001.

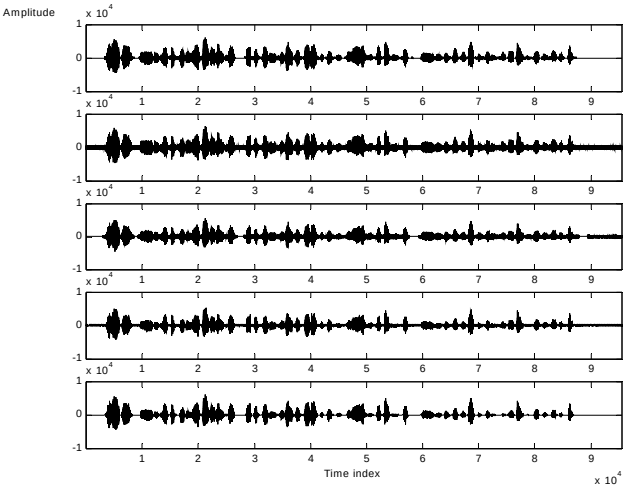

Fig. 3. Example of waveform plot, from top to bottom are the clean speech, noisy speech (corrupted by white Gaussian noise with segmental SNR equals 0 dB, enhanced speech using PSS, enhanced speech using MUWCT and enhanced speech using the proposed method.