

# Speech Enhancement using Weighting Function based on the Variance of Wavelet Coefficients

Ching-Ta Lu<sup>1,2</sup> and Hsiao-Chuan Wang<sup>1</sup>

Department of Electrical Engineering  
National Tsing Hua University, Taiwan<sup>1</sup>  
Department of Electronic Engineering  
Chin-Min College, Taiwan<sup>2</sup>

[Lucas1@ms26.hinet.net](mailto:Lucas1@ms26.hinet.net) and [Hcwang@ee.nthu.edu.tw](mailto:Hcwang@ee.nthu.edu.tw)

## Abstract

There are few works on the problem of heavy noise corruption in wavelet-based speech enhancement. In this paper, a new method is introduced to adapt the weighting function for wavelet coefficients (WCs) in each subband. The idea is based on that the change of WC variance in speech-dominated frames is larger than the change of WC variance in noise-dominated frames. We can define a weighting function for WCs in each subband so that WCs are preserved in speech-dominated frames and reduced in noise-dominated frames. Then a weighting function in terms of WC's variance is derived. The experimental results show that the proposed method is more robust than that of SNR adjusted speech enhancement system.

## 1. Introduction

In mobile communication, noise corruption is a serious problem in speech signal processing. Many studies had been made to reduce background noise of speech signal. The methods developed for this purpose include the spectral subtraction, the signal separation, and the wavelet-based approach. Especially the method based on wavelet transform has been extensively studied and proved to be a very promising one. Many of them deal with the cases of high SNR. There are few works on the problem of heavy noise corruption in wavelet-based speech enhancement.

Usually, the wavelet-based methods are developed for removing white Gaussian noise corruption [1]. Chang et al [2] proposed the node dependent thresholding for the adaptation of WCs in colored or non-stationary noise. They suggested a noise estimation method based on spectral entropy using histogram of intensity instead of median absolute deviation. Furthermore, they use a modified hard thresholding to alleviate time-frequency discontinuities. The performance is better than spectral subtraction and level dependent thresholding. Lu and Wang [3] suggested the adaptation of wavelet coefficient thresholds (WCTs) of each subband by using both the segmental signal to noise ratio (SegSNR) and noise masking thresholds (NMTs). The experiments show that the integration of SegSNR and NMTs can efficiently remove the background noise and suppress the residual noise. Furthermore, they suggested the method of adapting the weighting function to WCs for each subband. It could track the speech variation in each subband, and efficiently eliminate the

corrupting noise and get rid of the musical residual noise [4]. Jabloun and Champagne [5] proposed a frequency to eigen-domain transformation which provided a way to calculate a perceptual bound for the residual noise. It yielded an improved result where better shaping of the residual noise was achieved. Quatieri and Dunn [6] proposed an adaptive approach to enhance speech signal. It was developed based on auditory spectral change. The essence of the method is a Wiener filter. The degree of stationarity was derived from a signal change measurement, based on an auditory spectrum that accentuated change in spectral bands. Arslan et al [7] proposed the use of SNR to modify Wiener filter. It provided a much higher mean opinion score (MOS) than that of spectral subtraction.

Based on the above discussion, the performance of speech enhancement can be improved by modifying the weights of each wavelet subband. Both SNR and spectral change have been employed to modify the weighting filter [5][7-8]. Herein we propose a novel scheme which employs the change of WC variance to adapt the weighting function. Since the variance of noise WCs is stationary, whereas the difference of variances between two successive frames is small. On the other hand, the WCs of speech vary quickly in successive frames. The variance of WCs in speech changes more rapidly than that in noise. If the change of WC variance is large, the corresponding signal tends to be speech-dominated. We modify the weighting function by reducing the factor contributed by noise. This makes the weighting function near to one. The WCs are reserved and the speech distortion can be reduced. On the contrary, if the change of WC variance is small, the signal tends to be noise-dominated. We modify the weighting function by increasing the factor contributed by noise. This allows the reduction of WCs so that more background noise is removed.

The rest of the paper is organized as follows. Section 2 describes the block diagram of proposed speech enhancement system. Section 3 derives the proposed 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 for each subband is

estimated. This weighting function is kept until next speech-pause frame is detected. The change of variance is calculated by subtracting the WC variances of current frame from previous frame. These changes are used to adapt the weighting function.

The weighting function is multiplied to noisy speech WCs in each subband. Finally, the enhanced WCs are inversely transformed back to reconstruct the enhanced speech  $\tilde{s}_m(n)$ . In order to suppress the musical residual noise, the spectra are smoothed between successive frames by post-processing block in Fig. 1.

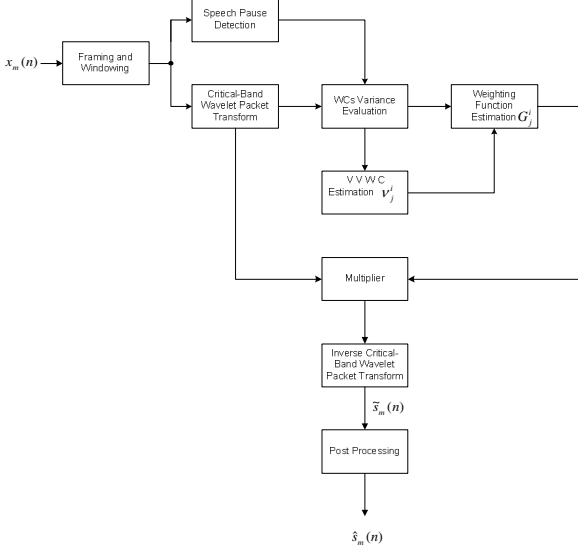


Figure 1: The procedure of proposed speech enhancement system.

### 3. The Proposed Weighting Function

For the case of additive noise corruption, the noisy speech signal can be modeled as the sum of clean speech and additive noise.

$$x_m(n) = s_m(n) + w_m(n) \quad n = 0, 1, 2 \dots N-1 \quad (1)$$

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.  $N$  is the number of samples in a frame.

Taking the wavelet transform, the noisy WCs can be expressed as the sum of speech WCs and noise WCs.

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

where  $X_{j,k}^i(m)$ ,  $S_{j,k}^i(m)$  and  $W_{j,k}^i(m)$  are noisy WC, speech WC, and noise WC at  $i$ -th subband in  $2^j$  scale, respectively.  $m$  represents the frame index.

#### 3.1 Weighting function for wavelet coefficients

By multiplying a weighting function  $G_j^i(m)$  to noisy WCs, we obtain

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

A mean-square error  $r_j^i(m)$  is defined for measuring the difference between speech WCs and enhanced WCs.

$$r_j^i(m) = \sum_k |\tilde{S}_{j,k}^i(m) - S_{j,k}^i(m)|^2 \quad (4)$$

The weighting factors  $G_j^i(m)$  are estimated by minimizing the mean-square error  $r_j^i(m)$ . We assume that  $s_m(n)$  and  $w_m(n)$  are mutually uncorrelated. Then the corresponding WCs are mutually uncorrelated also. The optimal weighting factor  $G_j^i(m)$  is derived as [4]

$$G_j^i(m) = \frac{\sigma_{S_j^i}^2(m)}{\sigma_{S_j^i}^2(m) + \sigma_{W_j^i}^2(m)} \quad (5)$$

where  $\sigma_{S_j^i}^2(m)$  and  $\sigma_{W_j^i}^2(m)$  denote the variance of speech WCs and noise WCs, respectively. The variance of noise WCs is evaluated in the speech-pause frame.

The variance of speech WCs is evaluated by subtracting the variance of noise WCs from that of noisy speech,

$$\sigma_{S_j^i}^2(m) = \max\{\sigma_{X_j^i}^2(m) - \sigma_{W_j^i}^2(m), 0\} \quad (6)$$

In order to improve the performance of removing background noise in speech-pause regions, the weighting function should be adapted by SegSNR. The weighting function Eq. (5) is modified to [4]

$$G_{j,SNR}^i(m) = \gamma(m) \cdot G_j^i(m) \quad (7)$$

where

$$\gamma(m) = \frac{1}{1 + e^{-a_1 \zeta(m) + b_1}} \quad (8)$$

$a_1$  and  $b_1$  are empirically chosen to be 0.2 and 2, respectively.  $\zeta(m)$  represents the posteriori SegSNR of  $m$ -th frame, and is defined by

$$\zeta(m) = 10 \cdot \log_{10} \left( \frac{\sum_n |x_n(m)|^2}{\sum_n |\tilde{w}_n(m)|^2} \right) \quad (9)$$

where  $\tilde{w}_n(m)$  denotes the estimated noise in a frame. It is updated when a speech-pause frame is detected and keeps unchanged until next speech-pause frame is detected.

#### 3.2 Modification of weighting function

The change of WC variance of noise is smaller than that of speech WCs. If the change of WC variance of a subband is large, the subband tends to be speech dominated. Decreasing the weighting to the variance of noise WCs in (5) makes the weighting function  $G_j^i(m)$  near to one. Most WCs of noisy speech are reserved. On the contrary, if the change of WC variance of a subband is small, the subband is noise-dominated. The weighting to the variance of noise WCs in (5) is increased. This results in the reduction of weighting function  $G_j^i(m)$ .

Therefore, the weighting function in Eq.(5) is suggested to be modified as

$$G_{j,vvwc}^i(m) = \frac{\sigma_{S_j^i}^2(m)}{\sigma_{S_j^i}^2(m) + \mu(m) \cdot \sigma_{W_j^i}^2(m)} \quad (10)$$

$\mu(m)$  is adapted by both SegSNR and the change of WC variance. It is formulated as

$$\mu(m) = \alpha_{SNR}(m) \cdot \rho(m) + [1 - \alpha_{SNR}(m)] \cdot \kappa_j^i(m) \quad (11)$$

where  $\rho(m)$  is a posteriori noise to signal ratio, and is given by

$$\rho(m) = \sum_n |\tilde{w}_n(m)|^2 / \sum_n |x_n(m)|^2 \quad (12)$$

The interpolating factor  $\alpha_{SNR}(m)$  in Eq.(11) is determined according to the posteriori noise to signal ratio  $\rho(m)$ , and can be written as

$$\alpha_{SNR}(m) = 1 - \frac{1}{1 + e^{-a_2 \cdot \rho(m) + b_2}} \quad (13)$$

In Eq. (11)  $\kappa_j^i(m)$  should be inverse proportional to the change of WC variance, and is defined as

$$\kappa_j^i(m) = c_5 \cdot [1 - |v_j^i(m) - \tilde{v}_j^i|] \quad (14)$$

where  $v_j^i(m)$  and  $\tilde{v}_j^i(m)$  represent the change of WC variance of  $m$ -th frame and the change of WC variance of estimated noise. The change of WC variance is defined by subtracting the variance of noisy WCs of previous frame from that of current frame given as

$$v_j^i(m) = |\sigma_{x_j^i}^2(m) - \sigma_{x_j^i}^2(m-1)| \quad (15)$$

In order to normalize the change of WC variance to be between zero and one, a maximal value in an utterance is used.

$$v_j^i(m) = v_j^i(m) / \max(v_j^i) \quad (16)$$

Avoiding the musical residual noise can be obtained by constraining the variation of weighting function in the successive frames. We slow down the variation of weighting for each subband by convolving the weighting function Eq.(10) with low-pass filter  $h$  for each subband.

$$G_{jvwc}^i(m) = h * G_{jvwc}^i(m) \quad (17)$$

An example of weighting function is demonstrated in Fig. 2. The weighting factor of each subband tends to be zero in the speech-pause regions.

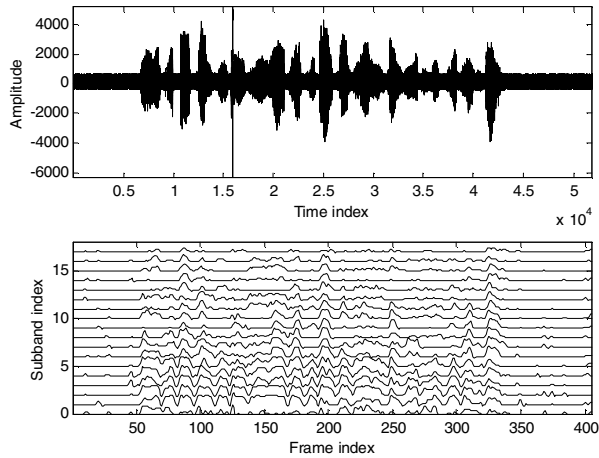


Figure 2: Example of weighting factors in each wavelet subband. The upper is the white Gaussian noise corrupted speech (SegSNR = 0dB). The underneath is the weighting factors, from top to bottom are subband 17 to subband 0, respectively.

#### 4. Experimental Results

In the experiments, the noisy speech is obtained by corrupting the clean speech with four kinds of noises. They are white Gaussian noise, babble noise, factory noise and F16-cockpit noise from Noisx-92 in three different noise levels. The noise levels are with segmental SNR equal to -10dB, -5dB and 0 dB. Test data are spoken in Mandarin with sampling frequency of 8 kHz, including five male and five female speakers.

Our proposed method (AW) is compared with the methods of modified Wiener filter (MWF) [7] and Teager energy based wavelet coefficient thresholding (TWCT) [9]. MWF is a generalized Wiener filter where a noise suppression factor is used. This noise suppression factor is updated according to frame-by-frame SNR to ensure the suppression on noise-only frames. TWCT is a level-dependent wavelet thresholding scheme which utilizes the Teager energy operator to improve the discriminative ability for determining whether a speech segment is speech-dominated or noise-dominated. The comparison of SegSNR improvement is shown in Table 1. The performance of our proposed method outperforms the other two methods, especially at the environments of very noisy cases. It reveals that the proposed algorithm benefits both low speech distortion and efficient noise reduction. Note that the above three speech enhancement methods use the same algorithm of speech-pause detection.

Table 1: Comparison of SegSNR improvement for the enhanced speech in various noises.

Noise type	SNR (dB)	Enhancement method		
		MWF	TWCT	AW
White Gaussian	-10	4.65	6.19	15.29
	-5	9.34	5.43	13.36
	0	9.23	4.54	10.08
F16	-10	3.07	6.86	12.03
	-5	3.72	6.06	10.42
	0	5.99	4.93	8.58
Factory	-10	2.74	6.97	11.00
	-5	2.82	6.03	9.13
	0	4.12	4.88	7.78
Babble	-10	1.64	5.63	7.28
	-5	1.65	5.35	7.31
	0	2.83	4.69	7.45

Fig. 3. illustrates the spectrograms of clean speech, noisy speech, and the enhanced speech obtained from various enhancement methods. The clean speech is corrupted with F16-cockpit noise with SegSNR equal to 0dB. The enhanced speech can be free from musical residual noise in the proposed scheme, whereas the background noise can be efficiently eliminated by the proposed method. Because the Wiener filter is derived with the assumption of the corrupting noise is white Gaussian, MWF can efficiently remove the white noise corrupted speech. However, it can not well cope with colored noise. The residual noise of enhanced speech for MWF method is the highest among the three algorithms. An informal listening test has been performed. The enhanced speech of the proposed scheme sounds more natural than the other two methods also.

Itakura-Saito (IS) distance is an objective quality measure that performs a comparison between spectral envelopes. This quality measure is more influenced by a mismatch in formant location than in spectral valleys [10]. The minimal value of IS distance corresponds to the best speech quality. Table 2. shows the performance of IS distance

for various speech enhancement methods. Observing the performance of MWF method, it can only well work at high SNR environments. It almost fails in case of with SegSNR lower than 0dB. Our method is proved to be efficient in removing the stationary noise.

### 5. Conclusions

A novel scheme based on the change of WC variance is presented. The algorithm of adapting the weighting function according to the change of WC variance can efficiently remove the background noise and suppress the residual noise in various noise levels. The experimental results show that our proposed method outperforms MWF method and TWCT wavelet-based methods.

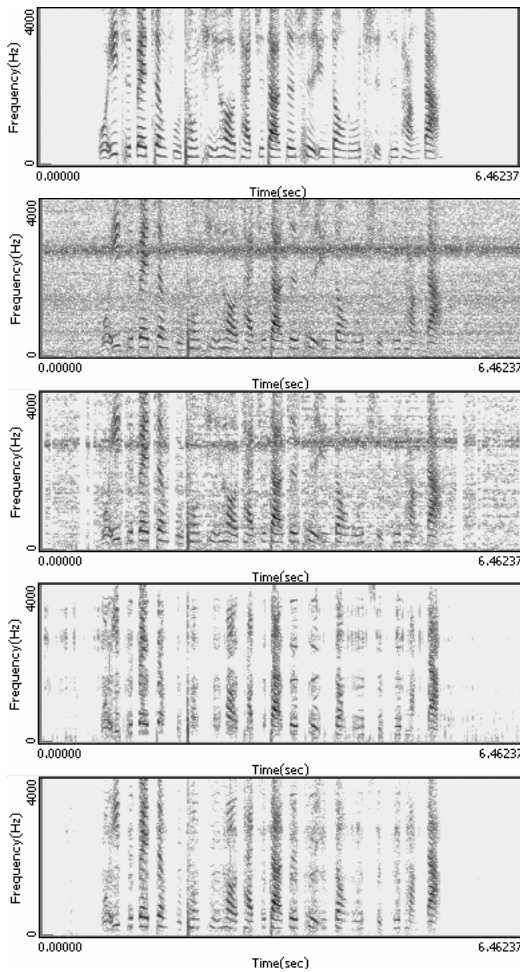


Figure 3: Spectrograms of clean speech (from top to bottom), noisy speech which corrupted by the F16-cockpit noise with SegSNR = 0 dB, enhanced speech using MWF method, TWCT method, and the proposed method AW, respectively.

### 6. Acknowledgements

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC--91-2219-E-007-017.

### 7. References

[1] Mallat, S., A Wavelet Tour of Signal Processing,

Academic Press, San Diego: A Harcourt Science and Technology, 1999.

[2] Chang, S., Kwon, Y., Yang, S. I. and Kim, I. J. "Speech Enhancement for Non-stationary Noise Environment by Adaptive Wavelet Packet," Proc. International Conference on Acoustic Signal Processing (ICASSP), pp. 561-564, May 2002.

[3] Lu, C. T. and Wang, H. C. "Enhancement of Single Channel Speech Using Perception-Based Wavelet Transform," Proc. International Conference on Spoken Language Processing (ICSLP), pp. Sep. 2002.

[4] Lu, C. T. and Wang, H. C. "Speech Enhancement Using Wavelet Transform with Constrained Thresholds," Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 81-84, Aug. 2002.

[5] Jabloun, F. and Champagne, B. "A Perceptual Signal Subspace Approach for Speech Enhancement in Colored Noise," Proc. International Conference on Acoustic Signal Processing (ICASSP), pp. 569-572, May 2002.

[6] Quantieri, T. F. and Dunn, R. B. "Speech Enhancement Based on Auditory Spectral Change," Proc. International Conference on Acoustic Signal Processing (ICASSP), pp. 257-260, May 2002.

[7] Arslan, L., McCree, A. and Viswanathan, V. "New Methods for Adaptive Noise Suppression," Proc. International Conference on Acoustic Signal Processing (ICASSP), pp. 812-815, May 1995.

[8] Martin, R., "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Proc.*, Vol. 9, NO. 5, pp. 504-512, 2001.

[9] 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. Jan. 2001.

[10] J. Deller, J. Prokis, and J. Hansen, *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

Table 2. Comparison of Itakura-Saito Distance for the enhanced speech in various noises.

Noise type	SNR (dB)	IS_Distance			
		Noisy	MWF	TWCT	AW
White Gaussian	-10	25.75	20.31	26.16	1.87
	-5	17.01	4.42	11.92	0.96
	0	8.31	1.44	3.56	0.72
F16	-10	2.06	1.81	2.70	1.20
	-5	1.58	1.37	1.71	0.72
	0	1.10	0.75	1.06	0.54
Factory	-10	1.92	2.20	2.64	1.44
	-5	1.48	1.56	1.51	0.91
	0	1.04	0.87	0.95	0.56
Babble	-10	1.85	2.07	2.46	1.86
	-5	1.45	1.62	1.52	1.07
	0	1.01	0.99	1.01	0.65