SPEECH ENHANCEMENT USING ROBUST WEIGHTING FACTORS FOR CRITICAL-BAND-WAVELET-PACKET TRANSFORM

Ching-Ta Lu[12] *and Hsiao-Chuan Wang*[1]

Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan[1]
Department of Electronic Engineering, Chin-Min College, Miaoli, Taiwan[2]
Lucas1@ms26.hinet.net and hcwang@ee.nthu.edu.tw

ABSTRACT

Although the noise masking threshold (NMT) has been applied in adapting the speech enhancement system, it may suffer from the underestimate in low-SNR condition. In this paper, we derive a robust weighting factor for each wavelet subband. The function of robust weighting factor is to keep the energy of residual noise lower than the NMT and the speech distortion smaller than the residual noise. If the energy of residual noise is greater than the NMT, the wavelet coefficients (WCs) of noisy speech are suppressed to remove more residual noise. If the energy of residual noise is smaller than the NMT, the weighting factor is set to one for retaining the speech quality. It results in a lower bound of NMT for preventing the underestimate of weighting factors. Experimental results show that the proposed method can improve the naturalness of enhanced speech.

1. INTRODUCTION

Wavelet analysis is able to track the tone in speech, and the subband decomposition can track the transient component of continuous speech signal. Therefore, the wavelet-packet transform (WPT) has been applied in speech enhancement progressively [1]. Many wavelet-based methods were developed for removing the white Gaussian noise [2]. They subtracted a threshold from wavelet coefficients (WCs) of noisy speech according to the noise level [3]-[6].

In case of corrupting by the colored noise, utilizing noise masking properties to adapt speech enhancement system is recently proposed [6-9]. Virag [7] utilized spectral subtraction algorithm to enhance noisy speech. Initially, the subtracted spectra are employed to estimate the noise masking threshold (NMT). Then the NMTs are incorporated to adjust the weighting factor of WCs for each subband. Hu and Loizou [8] proposed employing the masking properties in reducing the perceptual effect of the residual noise. The scheme was incorporated into a frequency-domain and a subspace-based method. Carnero and Drygajlo [6] formulated temporal and spectral psychoacoustic models of masking adapted to wavelet packet analysis. The algorithm of the proposed orthogonal wavelet packet transform permits them to efficiently approximate the auditory-critical-band decomposition in the time and frequency domains.

Based on the above findings, employing NMT to adapt speech enhancement algorithm is able to cope with noisy speech corrupted by colored noise. The performance of enhanced speech is characterized by a tradeoff between the amount of noise reduction, the speech distortion, and the level of musical residual noise [7]. In a high SNR environment, the estimated-speech spectra are accurate, and the estimated NMT is reliable. On the contrary, the background noise is generally overestimated for low SNR. It makes speech spectra to be underestimated, and so does NMT. Thus adequately limiting the lower bound of NMT is necessary. Despite the reliability of estimated NMT, conventional methods directly used the NMT to adapt the parameters of the enhancement system.

In this paper, two constraints are utilized to optimize the weighting factor of each subband. The first one is to keep the energy of residual noise lower than the NMT. If the energy of residual noise is greater than the NMT, the weighting factor becomes small to suppress the noise. On the contrary, if the energy of residual noise is smaller than the NMT, the noise can not be perceived by human ears. For retaining the speech quality, we do not need to change the WCs, i.e., the weighting factor is set to one. The second constraint guarantees that the speech distortion is smaller than residual noise. The experimental results show that our proposed method makes the enhanced speech more natural than that of NMT-only (named as perceptual weighting) adapted method.

2. DERIVATION OF WEIGHTING FACTORS

2.1. Problem formulation

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

$$x_m(n) = s_m(n) + w_m(n) \quad (1)$$

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.

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

$$X^i_{j,k}(m) = S^i_{j,k}(m) + W^i_{j,k}(m) \quad (2)$$

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

The WCs of enhanced speech are obtained by multiplying a weighting factor with WCs of noisy speech for each subband. It can be written as

$$\tilde{S}^i_{j,k}(m) = G^i_j(m) \cdot X^i_{j,k}(m) \quad (3)$$

The WC deviation (WCD) is defined as the difference between WCs of clean speech and enhanced speech.

$$E^i_{j,k}(m) = \tilde{S}^i_{j,k}(m) - S^i_{j,k}(m) \quad (4)$$

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

$$E^i_{j,k}(m) = G^i_j(m) \cdot X^i_{j,k}(m) - S^i_{j,k}(m) \quad (5)$$

Substitute (2) into (5), the WCD can be rewritten as

$$E^i_{j,k}(m) = S^i_{j,k}(m) \cdot [G^i_j(m) - 1] + G^i_j(m) \cdot W^i_{j,k}(m) \quad (6)$$

Therefore, WCD can be viewed as the sum of two parts,

$$E^i_{j,k}(m) = E_{S^i_{j,k}}(m) + E_{W^i_{j,k}}(m) \quad (7)$$

Since the contents of $s_m(n)$, $g_m(n)$ and $w_m(n)$ are all real numbers, the conjugate operator are omitted hereafter. The WC power of speech distortion and residual noise in 2^j scale are given as

$$P\{E_{S^i_j}(m)\} = \frac{1}{N_j}\sum_{k=0}^{2^{-j}-1} |S^i_{j,k}(m)|^2 \cdot |G^i_j(m) - 1|^2 \quad (8)$$

$$P\{E_{W^i_j}(m)\} = \frac{1}{N_j}\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2 \cdot |G^i_j(m)|^2 \quad (9)$$

Due to the mean of WCs is zero, and assuming the WCs sequence of signal and noise are mutually uncorrelated in a subband. The power of WCD can be expressed as the sum of (8) and (9).

2.2. Perceptual constraint

The objective is to minimize the power of speech distortion in WC with the constraint that the power of residual noise in WC keeps below a given threshold $\sigma^{i\,2}_j(m)$.

$$\min_{G^i_j(m)} \{P\{E_{S^i_j}(m)\}\} \quad \text{with the constraint } P\{E_{W^i_j}(m)\} \le \sigma^{i\,2}_j(m) \quad (10)$$

A speech-pause-detection algorithm is performed using the energy-based method. If a speech-pause frame is detected, the WCs are treated as the WCs of estimated noise. The noise WCs keep unchanged until next speech-pause frame is detected.

We propose that a threshold $\sigma^{i\,2}_j(m)$ is determined according to both the NMT and the ratio of speech distortion to residual noise in WC power. If the energy of residual noise is smaller than the NMT, the residual noise is not perceived by human ears. Accordingly, the WCs would be reserved to reduce the speech distortion. On the contrary, the WC of noisy speech should be suppressed if the noise level is greater than the NMT. Thereby the cost function $J(m)$ is formulated according to speech distortion and residual noise in WC power.

$$J^i_j(m) = P\{E_{S^i_j}(m)\} + \mu^i_j(m) \cdot [P\{E_{W^i_j}(m)\} - \sigma^{i\,2}_j(m)] \quad (11)$$

where $\mu^i_j(m)$ is the Lagrangian multiplier of a subband. It will be zero if the level of residual noise is under a given threshold.

Substitute (8) and (9) into (11), the cost function becomes

$$J^i_j(m) = \frac{1}{N_j}\sum_{k=0}^{2^{-j}-1} |S^i_{j,k}(m)|^2 \cdot |G^i_j(m) - 1|^2 + \mu^i_j(m) \cdot [\frac{1}{N_j}\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2 \cdot |G^i_j(m)|^2 - \sigma^{i\,2}_j(m)] \quad (12)$$

In order to minimize the cost function, (12) is partially differentiated with respect to the weighting factor $G^i_j(m)$, and set the differentiated result to zero. The optimized weighting factor can be calculated by the following formula,

$$G^i_j(m) = \frac{\sum_{k=0}^{2^{-j}-1} |S^i_{j,k}(m)|^2}{\sum_{k=0}^{2^{-j}-1} \{|S^i_{j,k}(m)|^2 + \mu^i_j(m) \cdot |W^i_{j,k}(m)|^2\}} \quad (13)$$

To differentiate (10) with respect to Lagrangian multiplier $\mu^i_j(m)$, and set the result to zero. The weighting factor becomes

$$G^i_j(m) = \sqrt{\frac{N_j \cdot \sigma^{i\,2}_j(m)}{\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2}} \quad , \; 0 \le G^i_j(m) \le 1 \quad (14)$$

Since (13) is equal to (14), the relation can be rearranged as

$$\sum_{k=0}^{2^{-j}-1} \{|S^i_{j,k}(m)|^2 + \mu^i_j(m) \cdot |W^i_{j,k}(m)|^2\} = \frac{\sum_{k=0}^{2^{-j}-1} |S^i_{j,k}(m)|^2 \cdot \sqrt{\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2}}{\sqrt{N_j} \cdot \sigma^i_j(m)} \quad (15)$$

The Lagrangian multiplier becomes

$$\mu^i_j(m) = \begin{cases} \frac{\sum_{k=0}^{2^{-j}-1} |S^i_{j,k}(m)|^2}{\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2} \cdot (\frac{\sqrt{\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2}}{\sqrt{N_j} \cdot \sigma^i_j(m)} - 1), & \text{if } \sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2 > N_j \sigma^{i\,2}_j(m) \\ 0, & \text{if } \sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2 \le N_j \sigma^{i\,2}_j(m) \end{cases} \quad (16)$$

A normalized threshold $N_j \cdot \sigma^{i\,2}_j(m)$ is chosen to be the NMT. The accuracy of NMT is very important to the performance of a system. Thus the threshold should be adequately constrained in a proper manner.

In the novel approach [8], the NMT is utilized to adapt the factor $\mu^i_j(m)$ in (13). The wrong estimation of NMT will deteriorate the performance of weighting factor. As discussion in the introduction, estimating a lower bound of NMT to prevent the underestimation of weighting factors is essential.

2.3 Robust constraint

In the robust criterion, constraining the speech distortion to be smaller than the residual noise is proposed, i.e.

$$\frac{P\{E_{S^i_{j,k}}(m)\}}{P\{E_{W^i_{j,k}}(m)\}} \le 1 \quad (17)$$

Substituting (8), and (9) into (17), we have

$$\frac{\sum_{k=0}^{2^{-j}-1} |S^i_{j,k}(m)|^2 \cdot |G^i_j(m) - 1|^2}{\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2 \cdot |G^i_j(m)|^2} \le 1 \quad (18)$$

Substituting weighting factor (13) into (18), it comes out

$$\mu^{i\,2}_j(m) \cdot \frac{\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2}{\sum_{k=0}^{2^{-j}-1} |S^i_{j,k}(m)|^2} \le 1 \quad (19)$$

Substituting the Lagrangian multiplier (16) into (19), we have

$$\left(\frac{\sum_{k=0}^{2^{-j}-1} |S^i_{j,k}(m)|^2}{\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2}\right)^2 \cdot \left(\frac{\sqrt{\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2}}{\sqrt{N_j} \cdot \sigma^i_j(m)} - 1\right)^2 \cdot \frac{\sum_{k=0}^{2^{-j}-1} |W^i_{j,k}(m)|^2}{\sum_{k=0}^{2^{-j}-1} |S^i_{j,k}(m)|^2} \le 1 \quad (20)$$

Decomposing (20), it can be rewritten as

$$\sum_{k=0}^{2^{-j}-1}|S_{j,k}^{i}(m)|^2\cdot\left(\frac{\sqrt{\sum_{k=0}^{2^{-j}-1}|W_{j,k}^{i}(m)|^2}-\sqrt{N_j}\cdot\sigma_j^{i}(m)}{\sqrt{N_j}\cdot\sigma_j^{i}(m)}\right)^2-\sum_{k=0}^{2^{-j}-1}|W_{j,k}^{i}(m)|^2\le 0 \quad (21)$$

The lower bound of the threshold $\sigma_j^{i\,2}(m)$ can be derived as

$$\sigma_j^{i}(m)\ge\sigma_{j\,\min}^{i}(m) \quad (22)$$

where

$$\sigma_{j\,\min}^{i}(m)=\frac{\sqrt{\sum_{k=0}^{2^{-j}-1}|S_{j,k}^{i}(m)|^2\cdot\sum_{k=0}^{2^{-j}-1}|W_{j,k}^{i}(m)|^2}}{\sqrt{N_j}\cdot\left(\sqrt{\sum_{k=0}^{2^{-j}-1}|S_{j,k}^{i}(m)|^2}+\sqrt{\sum_{k=0}^{2^{-j}-1}|W_{j,k}^{i}(m)|^2}\right)} \quad (23)$$

WC energy of clean speech, $\sum_{k=0}^{2^{-j}-1}|S_{j,k}^{i}(m)|^2$, is estimated by subtracting the WC energy of estimated noise from the WC energy of noisy speech.

3. THE PROPOSED ALGORITHM

Observing the lower bound in (23), it is adjusted by the magnitude between speech and noise in WC energy. Due to the robust criterion, the square root of the threshold should be greater than the lower bound. Integrating the noise masking property and robust weighting, NMT is the preliminary candidate of the given normalized threshold.

NMT can be estimated as follows. At first, the frequency analysis of a critical band is performed and a spreading function is applied to consider masking between different critical bands. Then the WC is subtracted by a threshold offset. Finally, the resulted WC is renormalized and compared with the absolute threshold of hearing. Detailed procedure can be found in [6-7][9]. The NMT must be greater than the normalized-lower bound, i.e.

$$N_j\cdot\sigma_{j\,\min}^{i\,2}(m)\le T_j^{i}(m) \quad (24)$$

From (13), we know that the Lagrangian multiplier should be not less than zero, hence (16) can be rewritten as

$$\mu_j^{i}(m)=\frac{\sum_{k=0}^{2^{-j}-1}|S_{j,k}^{i}(m)|^2}{\sum_{k=0}^{2^{-j}-1}|W_{j,k}^{i}(m)|^2}\cdot\max\left\{\frac{\sqrt{\sum_{k=0}^{2^{-j}-1}|W_{j,k}^{i}(m)|^2}}{\sqrt{N_j}\cdot\sigma_j^{i}(m)}-1\ ,\ 0\right\} \quad (25)$$

The relation between the proposed threshold and the NMT is summarized as

$$\sigma_j^{i\,2}(m)=\max\left\{\frac{T_j^{i}(m)}{N_j},\sigma_{j\,\min}^{i\,2}(m)\right\} \quad (26)$$

Substitute (25) into (13), the weighting factor can be simplified as

$$G_j^{i}(m)=1\Big/\left(1+\max\left\{\frac{\sqrt{\sum_{k=0}^{2^{-j}-1}|W_{j,k}^{i}(m)|^2}}{\sqrt{N_j}\cdot\sigma_j^{i}(m)}-1\ ,\ 0\right\}\right) \quad (27)$$

where $\sigma_j^{i}(m)$ can be obtained from (26).

4. EXPERIMENTS

In the experiment, the noisy speech is obtained by corrupting the clean speech with white Gaussian noise, F16-cockpit noise, factory noise, babble (speech-like) noise, and car noise with different noise levels. The speech-pause is detected in noisy environments. The performance is evaluated in average segmental SNR improvement and Itakura Saito (IS) distance in speech-activity regions.

The proposed method (Robust) is compared with perceptual weighting method (Perceptual). The structure of wavelet filterbank of critical-band-wavelet-packet transform can be found in [9]. Table 1 shows the segmental SNR (SegSNR) improvement with segment size equal to 256 for various noise levels. The performance of the robust weighting almost outperforms that of the perceptual weighting. Table 2 demonstrates the comparison of IS distance. It reveals that the proposed algorithm benefits low speech distortion and retains the residual noise in an acceptable level.

The waveforms of the enhanced speech are demonstrated in Fig. 1. Speech signal is corrupted by babble noise with SegSNR equaling 5 dB. It shows that both methods can efficiently remove the background noise. But the waveform of enhanced speech for perceptual weighting seems over-attenuated in low-SNR regions. Fig. 2 shows the comparison of spectrograms. The background noise can be efficiently removed by both methods. But the spectrum structure of based on robust weighting is finer than that of based on perceptual weighting. Thus the enhanced speech using our proposed method sounds more natural than that using perceptual weighting.

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

		SNR Improvement (dB)	
Noise type	SNR(dB)	Perceptual	Robust
White	0	4.01	5.12
	5	1.66	2.23
	10	-3.20	0.36
F16	0	3.03	3.33
	5	1.56	4.87
	10	-3.42	-0.02
Factory	0	2.90	3.02
	5	-0.18	2.23
	10	-3.35	-0.02
Babble	0	3.03	2.20
	5	-0.05	1.44
	10	-3.34	0.36
Car	0	3.75	2.52
	5	4.48	6.16
	10	-0.28	1.45

5. CONCLUSIONS

Our proposed method has been shown to be able to remove the background noise and perceptually suppress the residual noise. The experiments reveal that employing the proposed constraints can prevent the enhanced speech from over-attenuation.

Therefore, the enhanced version sounds more natural than that of NMT-only adapted algorithm.

ACKNOWLEDGEMENT

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC-92-2213-E-007-036.

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

Noise type	SNR (dB)	IS_Distance		
		Noisy speech	Perceptual	Robust
White	0	5.35	3.23	3.25
	5	3.59	2.98	1.61
	10	2.18	2.67	1.57
F16	0	2.60	2.76	2.28
	5	1.74	2.08	1.63
	10	1.02	2.05	1.17
Factory	0	2.36	2.66	2.38
	5	1.53	2.30	1.63
	10	0.88	2.00	1.17
Babble	0	2.08	2.33	2.12
	5	1.35	2.06	1.48
	10	1.76	1.71	1.57
Car	0	0.49	1.06	0.56
	5	0.10	1.07	0.46
	10	0.10	1.03	0.44

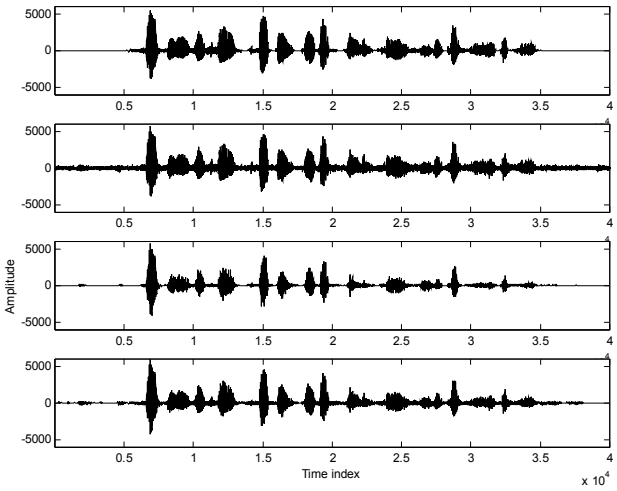

Fig. 1. Example of waveform plots (from top to bottom): the clean speech, the noisy speech (corrupted by babble noise with SegSNR = 5dB), and the enhanced speech using perceptual weighting method and robust weighting method.

REFERENCES

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

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

[3] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 613-627, May 1995.

[4] M. Jansen and A. Bultheel, "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.

[5] H. Sheikhzadeh and H. R. Abutalebi, "An improved wavelet-based speech enhancement system," *in. Proc. EuroSpeech*, 2001, pp. 1855-1858.

[6] B. Carnero and A. Drygajlo, "Perceptual Speech Coding and Enhancement Using Frame-Synchronized Fast Wavelet Packet Transform Algorithms," *IEEE Trans. Signal Processing*, vol. 47, pp. 1622-1635, June 1999.

[7] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 126-137, Mar. 1999.

[8] Y. Hu, and P.C. Loizou "A Perceptually Motivated Approach for Speech Enhancement," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 457-465, Sep. 2003.

[9] C. T. Lu, and H. C. Wang " "Enhancement of Single Channel Speech Based on Masking Property and Wavelet Transform," *Speech Commun*., vol. 41/2-3, p.p.409-427, 2003.

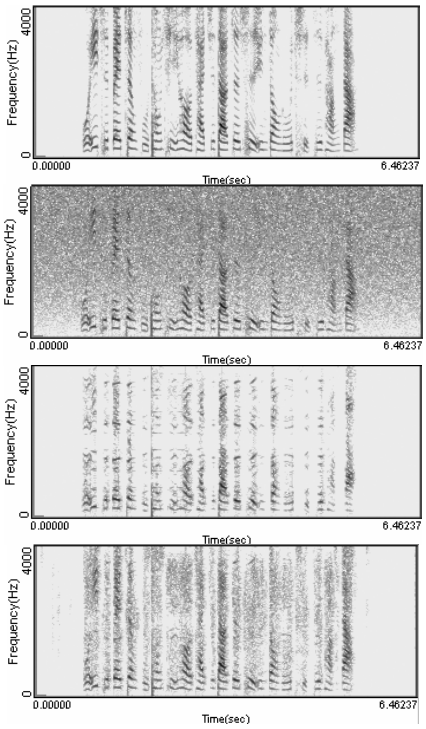

Fig. 2. Spectrograms of clean speech (from top to bottom): the clean speech, the noisy speech which corrupted by white noise with SegSNR = 5 dB, and the enhanced speech using perceptual weighting (Perceptual) method and proposed method (Robust).