A Time-Frequency Adaptation Based on Quantum Neural Networks for Speech Enhancement

Kun-Ching Wang[1] and Chiun-Li Chin[2]
[1]Department of Information Technology & Communication, Shin Chien University,
No. 200, University Rd, Neimen Shiang, Kaohsiung 845, Taiwan, R.O.C
wkc0224@seed.net.tw
[2]Department of Applied Information Sciences, Chung Shan Medical University Taichung, Taiwan, R.O.C

Abstract: In this paper, we propose a novel wavelet coefficient threshold (WCT) depended on both time and frequency information for providing robustness to non-stationary and correlated noisy environments. A perceptual wavelet filter-bank (PWFB) is firstly used to decompose the noisy speech signal into critical bands according to critical bands of psycho-acoustic model of human auditory system. The estimation of wavelet coefficient threshold (WCT) is then adjusted with the posterior SNR, which is determined by estimated noise power, through the well-known "Quantum Neural Networks (QNN)". In order to suppress the appearance of musical residual noise produced by thresholding process, we consider masking properties of human auditory system to reduce the effect of musical residual noise. Simulation results showed that the proposed system is capable of reducing noise with little speech degradation and the overall performance is superior to several competitive methods.

Key-Words: Speech enhancement; perceptual wavelet packet transformation; adaptive wavelet coefficient threshold; musical residual noise.

1 Introduction

Speech enhancement has become an important problem since there are many areas where it is necessary to enhance the perceptual quality of speech degraded by background noise, such as voice communication and coding systems, car interiors for hands free cellular, aircraft cockpits, hearing aids and automatic speech recognition (ASR) systems [1-2]. So far, the researchers provide many approaches to enhance speech quality [3-10]. Wavelet thresholding is a simple de-noise technique that adequately chooses the value of wavelet coefficient threshold (WCT) to remove noise form signal in many signal-processing applications [7-10]. Donoho and Johnston [7-8] proposed a universal threshold for removing the additive white Gaussian noise, but may not work well in enhancing colored-noise corrupted signal. After that, adaptive wavelet-based methods in speech enhancement are widely presented [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 the use of Teager energy operator (TEO) to improve the discriminability for a speech frame whether is speech-dominated or noise-dominated.

In fact, the key issue is to select threshold and shrinkage function in wavelet thresholding method. Traditional wavelet de-noising methods involve either hard or soft thresholding. In hard thresholding method, the coefficient is set to a specific value when its magnitude exceeds the threshold. On the other hand, soft thresholding shrinks or scales the coefficient that exceeds the threshold value. It is known that Quantum Neural Networks (QNN) is exploited a new shrinkage function in speech enhancement system [12]. To reduce the effect of musical residual noise, a number of methods were considered [13-14]. Virag [13] 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).

In this paper, we propose a novel WCT depended on both time and frequency information for providing robustness to non-stationary and correlated noisy environments. A perceptual wavelet filter-bank (PWFB) is firstly used to decompose the noisy speech signal into critical bands according to critical bands of psycho-acoustic model of human auditory system. The estimation of WCT is then adjusted with the posterior SNR, which is determined by estimated noise power, through the well-known QNN. In order to suppress the appearance of musical residual noise produced by thresholding process, we consider masking properties of human auditory system to reduce the effect of musical residual noise.

2 Perceptual Wavelet Filter-bank

The human speech mostly spans within 4 kHz and there are only 17 critical bands existed in this bandwidth as listed in Table 1 [13]. In order to introduce a speech enhancement method based on the human auditory model, a perceptual wavelet filter-bank (PWFB) is designed to mimic the critical bands as widely used in perceptual [15]. The filter banks, implemented by using the high-pass filter and low-pass filter with the Daubechies family wavelet [16], perceptually divide whole band into subband domain. In the first level decomposition, scaling space and wavelet space will be decomposed into two which correspond to the frequency ranges of 0–2 and 2–4 kHz. This operation is repeated to at most five times. Table 2 shows the coefficients through five-stage tree structure of perceptual wavelet filter-bank (PWFB). The PWFB decomposes the noisy signal $x(n)$ into 17-subbands corresponding to wavelet coefficient sets

$$w_c^{k,m} = PWFB\{x(n)\},\ \ n = 1,....,N, \tag{1}$$

where $PWFB\{\cdot\}$ means the perceptual wavelet packet transform. $w_c^{k,m}$ defines the cth wavelet coefficient in kth subband. N is the length of speech frame.

3 The Estimation of Time-Frequency Dependent WCT

Here we use a new adaptive time-frequency dependent thresholds estimation method. This involves first estimating the standard deviation of the noise, σ, for every subband and time frame. Consequently, we use a quantile-based noise tracking approach to track the slowly varying non-stationary noise statistics [17].

Table 1. The characteristics of critical bands under 4 kHz

Bark-Band Number	Lower Edge (Hz)	Upper Edge (Hz)	Center frequency (Hz)	Bandwidth (Hz)
1	0	100	50	100
2	100	200	150	100
3	200	300	250	100
4	300	400	350	100
5	400	510	450	110
6	510	630	570	120
7	630	770	700	140
8	770	920	840	150
9	920	1080	1000	160
10	1080	1270	1170	190
11	1270	1480	1370	210
12	1480	1720	1600	240
13	1720	2000	1850	280
14	2000	2320	2150	320
15	2320	2700	2500	380
16	2700	3150	2900	450
17	3150	3700	3400	550

Table 2. The coefficients of perceptual wavelet filter-banks

Bark-Band Number	Transform stage	Coefficients index	Coefficients length
1	5	0-7	8
2	5	8-15	8
3	5	16-23	8
4	5	24-31	8
5	5	32-39	8
6	5	40-47	8
7	5	48-55	8
8	5	56-63	8
9	4	64-79	16
10	4	80-95	16
11	4	96-111	16
12	4	112-127	16
13	4	128-143	16
14	4	144-159	16
15	3	160-191	32
16	3	192-223	32
17	3	224-255	32

The noise level estimation is given by

$$\tilde{\sigma}^{k,m} = \sqrt{\sum_{j=0}^{\text{int}(q\cdot L_{seg}^k)} \left|w_c^{k,m}\right|^2 \Big/ \text{int}(q\cdot L_{seg}^k)}, \tag{2}$$

where int(·) is the nearest integer rounding function. The nominal value of q is 0.2. $\tilde{\sigma}^{k,m}$ is denoted as the corresponding estimated noise level of the mth frame in the kth subband. These are estimated using the segment of previous data $\left\{w_c^{k,m} \mid c = 0,...,L_{seq}^k - 1\right\}$.

where L_{seq}^k means the length of the segment in the kth subband.

The initial WCT, $WCT_0^{k,m}$, for kth subband at the mth frame, can be estimated as in [17]:

$$WCT_0^{k,m} = \tilde{\sigma}^{k,m} \cdot \sqrt{2\log(L_{frm}^k \log_2(L_{frm}^k))}, \tag{3}$$

where L_{frm}^k means the frame length at the kth subband.

The posteriori SNR on kth subband can be evaluated as

$$SNR_{pot}^{k,m} = 10\cdot\log_{10}\left(\frac{\left|w_c^{k,m}\right|^2}{\left|\tilde{\sigma}^{k,m}\right|^2}\right), \tag{4}$$

where $\left|\tilde{\sigma}^{k,m}\right|^2$ and $\left|w_c^{k,m}\right|^2$ denote the estimated subband noise power and observed signal power.

To remedy the drawbacks of the traditional threshold algorithms, we adopt a new method of S-curve of Quantum Neural Network (QNN) to choice appropriate WCT [18]. The S-curse in QNN model is multi-level, which expression is

$$QNN(x) = \frac{1}{n_s}\sum_{i=1}^{n_s}\frac{1}{1+\exp(-(x-\theta_i))}, \tag{5}$$

where n_s is the number of curve level and θ_i is the position of the level respectively. x is herein defined as $x = \alpha \cdot (SNR_{pot}^{k,m} - T)$.

4 Perceptual Suppression using Noise Masking Threshold (NMT)

In order to improve the final perceptual quality, a suppression method of musical residual noise can adopt a perceptual gain factor into wavelet thresholding. The time-frequency adapted wavelet threshold is finally modified as below:

$$WCT_{final}^{k,m} = WCT_0^{k,m} \cdot Gain_{pecp}^{k,m},$$

where (6)

$$Gain_{pecp}^{k,m} = 1 \Big/ \left(1 + \max \left\{ \sqrt{\frac{\left|\tilde{\sigma}^{k,m}\right|^2}{NMT(k,m)}} - 1,\ 0 \right\} \right)$$

denotes a perceptual gain factor given by [17].

From Eq.(6), it is known that if the energy of musical residual noise, $\left|\tilde{\sigma}^{k,m}\right|^2$, is greater than the NMT, $NMT(k,m)$, 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.

In order to calculate the NMT on each subband, the estimated spectra of enhanced speech must be first determined and it be roughly estimated by the spectral-subtraction method. Netx, the subband energy $\varepsilon(k,m)$ is calculated by

$$\varepsilon(k,m) = \sum_{k_l}^{k_h} \left| w^{k,m} \right|^2, \quad (7)$$

where h_k and l_k denote the upper and the lower frequencies at critical band can be found in [13].

An excitation pattern $B(k,m)$ can be regarded as an energy distribution along the basilar membrane. $B(k,m)$ can be calculated by convolving the subband energy $\varepsilon(k,m)$ with the spreading function $F(k)$. $B(k,m)$ is given by [13, 19]:

$$B(k,m) = F(k) * \varepsilon(k,m). \quad (8)$$

A relative threshold offset $O(k)$, which can be found in [13, 19], 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 $\tilde{B}(k,m)$ is computed as the sum of the log energy for the excitation pattern and the offset $O(k)$, written as

$$\tilde{B}(k,m) = 10 \cdot \log_{10} B(k,m) + O(k), \quad (9)$$

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

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

$$Th(k,m) = \tilde{B}(k,m) - G(k,m), \quad (10)$$

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

$G(k,m)$ is expressed as

$$G(k,m) = 10 \cdot \log_{10} \left(\frac{B(k,m)}{\varepsilon(k,m)} \right). \quad (11)$$

Additionally, the normalized threshold $Th(k,m)$ is compared with the absolute-hearing threshold (AHT) which is frequency-dependent and can be closely approximated as [13, 19],

$$\mathrm{AHT}(f) = 3.64 f^{-0.8} - 6.5 e^{-0.6(f-3.3)^2} + 0.001 f^4 \quad \text{[dB]} \quad (12)$$

with f in kilohertz.

Finally, the NMT $NMT(k,m)$ is obtained by

$$NMT(k,m) = \max\left\{ \mathrm{AHT}(f),\ Th(k,m) \right\}, \quad (13)$$

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

5 Implementation of the Proposed Algorithm

Figure 2 shows the system block diagram for the proposed wavelet-based speech enhancement algorithm. The wavelet packet filter-bank is first applied to decompose the noisy speech signal into multi-resolution time-spectral subbands. Thresholds are independently estimated across successive speech frames in each decomposed subband, and are adapted as time-variant values based on the adaptive noise estimation algorithm. The suppression of background noise is then achieved by soft thresholding the decomposed wavelet coefficients. Finally, these thresholded wavelet coefficients are reconstructed to obtain the enhanced speech samples using the inverse wavelet packet filter-bank.

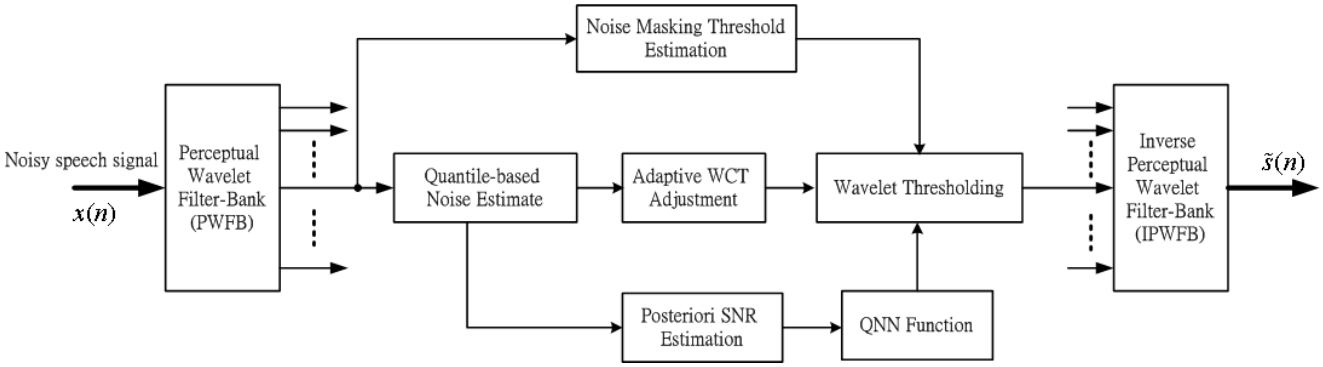

Figure 2. The architecture of proposed speech enhancement method based on the time-frequency adaptation of the wavelet threshold

Table 3. The average SegSNR results of speech enhancement under various noisy conditions

Noise type	average SegSNR (dB)		
	Proposed	*S.H. Chen*	*M. Bahoura*
F16 noise	***10.62***	10.12	7.81
White noise	***9.57***	9.34	6.98
Babble noise	***7.74***	7.40	5.37
Factory noise	***8.65***	9.12	7.12
Vehicle noise	***12.65***	10.57	8.24

Table 4. MOS results of the listening test

Noise type	The mean opinion score		
	Proposed	*S.H. Chen*	*M. Bahoura*
F16 noise	***3.57***	3.54	*2.12*
White noise	***3.74***	3.12	*2.89*
Babble noise	***2.56***	2.67	*1.68*
Factory noise	***3.13***	2.89	*1.45*
Vehicle noise	***3.89***	3.16	*2.45*

6 Experimental Results

In this experiments, the speech databases are Mandarin and spoken by 22 males and 10 females. The frame size is 256 at the sampling rate of 8 kHz with 16-bit resolution. Noisy speech signals were obtained by corrupted the clean speech with White, F16 and Babble (speech-like) noises extracted from the Noisex-92 database [20]. The recorded speech signal was tested in the noisy environment including SNRs range from - 5dB to 10 dB. In our experiments, the subjective evaluation and objective evaluation are applied to evaluate the performance of the speech enhancement method.

3.1 Segment SNR Improvement

The average SegSNR can be used to estimate the amount of noise reduction, residual and speech distortion. Table 3 shows that the average SegSNR results of the speech enhancement evaluations in different SNR levels. From this Table we can see that the proposed algorithm has much better enhancement performance than others.

3.2 Subjective Listening Tests

The mean opinion score (MOS) [21] was used to represent the global perception of the residual noise, background noise and speech distortion. The MOS is subjectively evaluated the subjective listening tests by a five-scale absolute opinion from 1 (poor) to 5 (excellent). The results of subjective listening tests are presented in Table 4. 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. The perceptual method is better able to remove the background noise than that without perceptual model.

7 Conclusion

In the paper, a novel speech enhancement algorithm using time-frequency wavelet threshold is presented. In order to exploit the physiology of human auditory system to recovery high-quality speech from noisy speech, the noisy speech is first decomposed into critical bands by perceptual wavelet packet transform. Then, an adaptive wavelet threshold is adjusted according to posterior SNR based on S-curve of Quantum Neural Network (QNN). Experimental results show that the proposed algorithm is better able to perceptually reduce the non-stationary and colored noise and is free from musical residual noise.

8 Acknowledgment

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, *in Proceedings of Int. Conf. Spoken Language Process '90*, 1990, pp. 1113-1116.

[2] J.H. Chen and A. Gersho, Adaptive Postfiltering for Quality Enhancement of Coded Speech, *IEEE Trans. Speech and Audio Processing*, Vol. 3, 1995, pp. 57-71.

[3] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoustics Speech Signal Process*, Vol. 27, 1979, pp.113-120.

[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, 1984, pp. 1109-1121.

[7] D.L. Donoho, De-noising by Soft Thresholding, *IEEE Trans. Inform. Theory*, Vol. 41, May, 1995, pp. 613-627.

[8] D.L. Donoho and I.M. Johnstone, Ideal Spatial Adaptation by Wavelet Shrinkage, *Biometrika*, Vol. 81, No. 3, 1994, pp. 425-455.

[9] S.H. Chen and J.F. Wang, Speech Enhancement Using Perceptual Wavelet Packet Decomposition and Teager Energy Operator, *Journal of VLSI Signal Processing*, Vol. 36, 2004, pp. 125-139.

[10] S.F. Lei and Y.K. Tung, Speech Enhancement for Nonstationary Noises by Wavelet Packet Transform and Adaptive Noise Estimation, *Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems*, Dec. 2005, pp. 41-44.

[11] M. Bahoura and J. Rouat, Wavelet speech enhancement based on the teager energy operator, *IEEE Signal Process. Lett.* Vol. 8, No. 1, 2001, pp. 10-12.

[12] L Fei, Z Shengmei and Z Baoyu, Quantum neural network in speech recognition - *Proc. IEEE Int. Conf. Signal Process*, 2000

[13] N. Virag, Single channel speech enhancement based on masking properties of the human auditory system, *IEEE Trans. Speech Audio Process.*, Vol. 7, No. 2, Mar. 1999, pp. 126-137.

[14] Y. Hu and P.C. Loizou, Speech enhancement based on wavelet thresholding the multitaper spectrum, *IEEE Trans. Speech Audio Process*, Vol. 12, No. 1, 2004, pp. 59-67.

[15] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, New York, 1990.

[16] S. Mallat, Multifrequency channel decomposition of images and wavelet model, *IEEE Trans. Acoust. Speech Signal Process*, Vol. 37, 1989, pp. 2091-2110.

[17] Q Fu and E.A. Wan, Perceptual wavelet adaptive denoising of speech, *Eighth European Conference on Speech*, 2003

[18] J. F. Teng, J. Dong, S. Y. Wang, H. Bao, M. G. Wang, A Speech Enhancement Algorithm Based on Bark-Scale Wavelet Package, *proceedings of the sixth international conference on machine learning and cybernetics*, Hong Kong, August 2007, pp.19-22.

[19] C.T. Lu and H.C. Wang, Speech enhancement using perceptually constrained gain factors in critical-band-wavelet packet transform, *IEE Electron. Lett.*, Vol. 40, No. 6, 2004, pp. 394-396.

[20] 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 Commun.*, Vol. 12, 1993, pp. 247-251.

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