SPEECH ENHANCEMENT BY WAVELET PACKET TRANSFORM WITH BEST FITTING REGRESSION LINE IN VARIOUS NOISE ENVIRONMENTS

Sung-il Jung†, Younghun Kwon‡ and Sung-il Yang†

† Department of Electrical and Computer Engineering, Hanyang University, Korea
‡ Department of Physics, Hanyang University, Korea
E-mail: sijung@ihanyang.ac.kr

ABSTRACT

In this paper, we suggest a speech enhancement method which can be applied in various noise environments. This method uses a wavelet packet transform (WPT) and a best fitting regression line (BFRL) in order to accurately estimate parameters for the spectral subtraction method based on the time-varying gain function. It should be noted that our method does not use the statistical information of pause region detected by voice activity detector. The evaluation is performed on various environments where the noisy speech are between SNR -5 ~ 15 dB, in various noises. We compare the performance of the proposed method, with that of magnitude spectral subtraction in WPT and nonlinear magnitude spectral subtraction in WPT. We can see that the performance of the proposed method is better than that of any other methods, with regard to objective test (segmental SNR, weighted spectral slope), spectrogram analysis, and subjective one (mean opinion score). Especially, our method showed reliable result even at low SNR.

1. INTRODUCTION

Spectral subtraction method is widely used on the single channel where speech coexists with noise. Usually, spectral subtraction method uses the statistical information of pause region, detected by voice activity detector (VAD) [1, 2, 4, 5]. However, if background noise is nonstationary or the speech is at low SNR, it will be difficult to use VAD. Consequently, in various noise-level conditions, the noise estimation using VAD will not be satisfactory. Also, the enhanced speech obtained by conventional spectral subtraction has shortcoming that the speech contains the perceptually annoying *musical tone* [1]. To reduce the musical tone, various approaches, based on spectral subtraction-type algorithm, have been proposed: for examples, magnitude averaging method [2], one with spectral flooring factor [3], one with oversubtraction based on masking property of the auditory system [1, 4], one using geometric mean [5], and one with optimal MMSE estimation of the short-time spectral amplitude [6], etc. However, any of these methods could not efficiently eliminate the musical tone in the enhanced speech.

In previous research, we applied a wavelet packet transform (WPT) [7, 8] to spectral subtraction method for speech enhancement. Even though we could enhance the noisy speech by the method, there still existed quite much musical tone in the enhanced speech. So, in order to figure out the problem, we suggest a speech enhancement method which uses not only the WPT but also the best fitting regression line (BFRL) [8]. In the new method, we introduce the nonlinear spectral subtraction with a modified time-varying gain function [5], consisting of an adaptive noise estimation and an adaptive oversubtraction factor using relative magnitude difference measure (RMDM), without the statistical information of pause region.

2. PROPOSED SPEECH ENHANCEMENT

Noisy speech $x(n)$ can be expressed as sum of additive noise $\omega(n)$ and clean speech $s(n)$.

$$x(n) = s(n) + \omega(n) \tag{1}$$

Coefficient of wavelet packet transformation (CWPT) for noisy speech $x(n)$ can be written as follows;

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

where i, j, k, and m are frame index, node index $(0 \le j \le 2^{K-k} - 1)$, tree depth $(0 \le k \le K)$ in total tree depth K, and the CWPT index in the node, respectively. $S_{i,j}^{k}(m)$ is the CWPT of clean speech and $W_{i,j}^{k}(m)$ that of noise.

Usually, it is very difficult to obtain accurately the parameters for the time-varying gain function from the CWPT $X_{i,j}^{k}(m)$. Thus, in order to optimally estimate these parameters, we use the BFRL $\overline{\mathbf{X}}_{i,j}^{k} = [\overline{X}_{i,j}^{k}(0), \cdots, \overline{X}_{i,j}^{k}(N-1)]^{\mathrm{T}}$

with the error expectation $E[\mathbf{e}_{i,j}^k]=0$ in equation (3), which can be obtained by least square method [8].

$$|\mathbf{X}_{i,j}^k| = \mathbf{A}\mathbf{c}_{i,j}^k + \mathbf{e}_{i,j}^k \tag{3}$$

$$\bar{\mathbf{X}}_{i,j}^k = \mathbf{A}\mathbf{c}_{i,j}^k = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}|\mathbf{X}_{i,j}^k| \tag{4}$$

where $|\mathbf{X}_{i,j}^k| = [|X_{i,j}^k(0)|,\cdots,|\bar{X}_{i,j}^k(N-1)|]^{\mathrm{T}}$ is the magnitude of coefficients in wavelet packet node (MCWPN) with uniform band structure. $\mathbf{e}_{i,j}^k = [e_{i,j}^k(0),\ldots,e_{i,j}^k(N-1)]^{\mathrm{T}}$ denotes the error and $\mathbf{A}$ is defined as the transform matrix of $N\times 2$.

At first, let us propose the modified adaptive noise estimation method, given by equation (5) ~ (7). This method uses the noise estimation of previous frame only, without the statistical information of pause region. We can estimate the amount of noise in the subband ($\mathrm{SB_{NS}}$) consisting of several nodes, by the magnitude SNR (MSNR), and evaluate the correlation degree between $\tilde{W}_{i-1,j}^k(m)$ and $\bar{X}_{i,j}^k(m)$ in the $\mathrm{SB_{NS}}$, by the correlation coefficient (CC). $\tilde{W}_{i,j}^k(m)$, MSNR and, CC can be defined as follows;

$$\tilde{W}_{i,j}^k(m) = \begin{cases} (1-\kappa\mathrm{MSNR}_i(\xi))\tilde{W}_{i-1,j}^k(m) + \kappa\mathrm{MSNR}_i(\xi)\bar{X}_{i,j}^k(m), \\ \quad \textbf{if } \mathrm{MSNR}_i(\xi) > \mathrm{Th_{MSNR}} \textbf{ AND } \mathrm{CC}_i(\xi) > \mathrm{Th_{CC}} \\ \tilde{W}_{i-1,j}^k(m), \textbf{ otherwise} \end{cases} \tag{5}$$

$$\mathrm{MSNR}_i(\xi) = \min\left(1, \frac{\sum_{m=\mathrm{SB_{NS}}\cdot\xi}^{\mathrm{SB_{NS}}\cdot(\xi+1)} \tilde{W}_{i-1,j}^k(m)}{\sum_{m=\mathrm{SB_{NS}}\cdot\xi}^{\mathrm{SB_{NS}}\cdot(\xi+1)} \bar{X}_{i,j}^k(m)}\right) \tag{6}$$

$$\mathrm{CC}_i(\xi) = \frac{\sum_{m=\mathrm{SB_{NS}}\cdot\xi}^{\mathrm{SB_{NS}}\cdot(\xi+1)} (\bar{X}_{i,j}^k(m)-\mu_{\bar{\mathbf{X}}_i}(\xi))(\tilde{W}_{i-1,j}^k(m)-\mu_{\tilde{W}_{i-1}}(\xi))}{\sigma_{\bar{\mathbf{X}}_i}(\xi)\sigma_{\tilde{W}_{i-1}}(\xi)} \tag{7}$$

where $\mathrm{SB_{NS}}$ is given by the multiplication between node size and node batch $2^p (k \le p)$ divided into node 2^{K-k} on k level of tree depth, and ξ $(0 \le \xi \le 2^{K-p}-1)$ is $\mathrm{SB_{NS}}$ index. κ $(\kappa < 1)$ denotes the additive weight of MSNR. $\mathrm{Th_{MSNR}}$ is the threshold of the MSNR and $\mathrm{Th_{CC}}$ that of the CC.

Next, we suggest the method to correctly evaluate the adaptive oversubtraction factor, by the RMDM. Equation (8) represents the RMDM $\gamma_i(\tau)$ per subband (SB_γ) consisting of several nodes. If $\gamma_i(\tau)$ is closed to 0 (1), this subband can be considered as noise-like one (speech-like one). SB_γ is employed in order to use the differential oversubtraction factor which has different value in each subband.

$$\begin{aligned} \gamma_i(\tau) &= \frac{2\sqrt{\sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} |X_{i,j}^k(m)| \sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} |W_{i,j}^k(m)|}}{\sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} |X_{i,j}^k(m)| + \sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} |W_{i,j}^k(m)|} \\ &= \frac{2\sqrt{\sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \bar{X}_{i,j}^k(m) \sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \bar{W}_{i,j}^k(m)}}{\sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \bar{X}_{i,j}^k(m) + \sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \bar{W}_{i,j}^k(m)} \\ &= \sqrt{1-\left(\frac{\sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \bar{S}_{i,j}^k(m)}{\sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \bar{X}_{i,j}^k(m) + \sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \bar{W}_{i,j}^k(m)}\right)^2} \end{aligned} \tag{8}$$

where SB_γ and SB_γ index τ are defined identically as $\mathrm{SB_{NS}}$ and $\mathrm{SB_{NS}}$ index ξ in equation (5) ~ (7). Because expectation $E[|\mathbf{X}_{i,j}^k|] = E[|\mathbf{S}_{i,j}^k|] + E[|\mathbf{W}_{i,j}^k|]$ of the MCWPN is equal to expectation $E[\bar{\mathbf{X}}_{i,j}^k] = E[\bar{\mathbf{S}}_{i,j}^k] + E[\bar{\mathbf{W}}_{i,j}^k]$ of the BFRL, the MCWPN can be rewritten in terms of the BFLR, in equation (8). Generally, it is difficult to estimate accurately $\bar{W}_{i,j}^k(m)$ from $\bar{X}_{i,j}^k(m)$. However, it is reasonable to regard $\bar{X}_{i,j}^k(m)$ of smaller value, than proposed noise estimation $\tilde{W}_{i,j}^k(m)$, as the noise. Then we calculate the $\gamma_i(\tau)$ using the $\bar{X}_{i,j}'^k(m)$ $(= \max(\bar{X}_{i,j}^k(m), \tilde{W}_{i,j}^k(m)))$.

$$\gamma_i(\tau) \cong \frac{2\sqrt{\sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \bar{X}_{i,j}'^k(m) \sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \tilde{W}_{i,j}^k(m)}}{\sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \bar{X}_{i,j}'^k(m) + \sum_{m=\mathrm{SB}_\gamma\cdot\tau}^{\mathrm{SB}_\gamma\cdot(\tau+1)} \tilde{W}_{i,j}^k(m)} \tag{9}$$

$$\psi_i(\tau) = \begin{cases} \rho\dfrac{\gamma_i(\tau)-\mathrm{Th_{RMDM}}}{1-\mathrm{Th_{RMDM}}}, \textbf{if } \gamma_i(\tau) > \mathrm{Th_{RMDM}} \\ 0, \qquad\qquad \textbf{otherwise} \end{cases} \tag{10}$$

where $\mathrm{Th_{RMDM}}$ is the threshold of RMDM and ρ the weight to decide the maximum oversubtraction factor. As shown in equation (10), the proposed adaptive oversubtraction factor $\psi_i(\tau)$ can be calculated by $\gamma_i(\tau)$

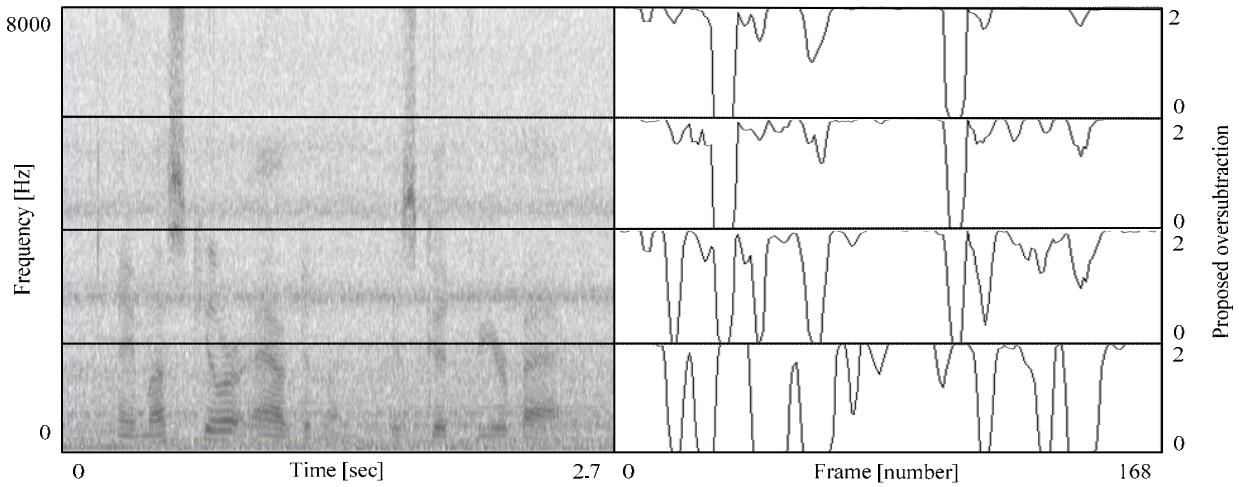

Figure 1: Speech spectrogram of noisy speech at SNR 5 dB, corrupted with aircraft cockpit noise and the diagrams of the proposed oversubtraction factor per subband

in equation (9). For example, if we set $\rho = 2$ and measure $\psi_i(\tau)$ from noisy speech at SNR 5 dB, corrupted by aircraft cockpit noise, we can obtain the result in figure 1. Especially, we can observe that $\psi_i(\tau)$ represents properly the characteristic of speech even in the very noisy speech region where we hardly extract the speech signals from the noise ones. Equation (11) represents the modified version of time-varying gain function. For speech enhancement, we apply equation (11) to equation (12).

$$G_{i,j}^k(m) = \begin{cases} \sqrt{1 - \dfrac{(1+\psi_i(\tau))\tilde{W}_{i,j}^k(m)}{\bar{X}_{i,j}^k(m)}}, & \textbf{if } \dfrac{\tilde{W}_{i,j}^k(m)}{\bar{X}_{i,j}^k(m)} < \dfrac{1}{1+\psi_i(\tau)} \\ \eta_i(\tau)\sqrt{\dfrac{\tilde{W}_{i,j}^k(m)}{\bar{X}_{i,j}^k(m)}}, & \textbf{otherwise} \end{cases} \tag{11}$$

$$\tilde{S}_{i,j}^k(m) = G_{i,j}^k(m) X_{i,j}^k(m) \tag{12}$$

where $\eta_i(\tau)$ $(0 \le \eta_i(\tau) < 1)$ is the spectral flooring factor and $\tilde{S}_{i,j}^k(m)$ the CWPT of enhanced speech.

3. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed speech enhancement method. For the WPT, we use the Daubechies basis with uniform band structure [7]. And we set $\mathrm{Th}_{\mathrm{MSNR}}$, $\mathrm{Th}_{\mathrm{CC}}$, κ in equation (5) to be 0.5, 0.75 and 0.25, respectively. We use ρ, $\mathrm{Th}_{\mathrm{RMDM}}$ in equation (10) and $\eta_i(\tau)$ in equation (11) as 2, $2\sqrt{2}/3$, and 0.001. For the experiments, we choose the 20 sentences from TIMIT, and 3 type of noises: aircraft cockpit noise, speech-like noise, and white Gaussian noise from NoiseX-92. Using these speech and noise, we generate 3 different noisy speech at SNR -5 ~ 15 dB, corrupted by 3 type of noises. The evaluation is performed in comparison with the magnitude spectral subtraction in wavelet packet transform (MSS_WPT), the nonlinear magnitude spectral subtraction in wavelet packet transform (NMSS_WPT), and the proposed method (PM). For the performance evaluation, we use the testing methods such as the objective test (segmental SNR [10], weighted spectral slope (WSS) [10]), speech spectrogram analysis [4], and subjective test (mean opinion score (MOS) [10]). When we perform the NMSS_WPT, we use the adaptive oversubtraction factor [4] between 1 ~ 5, based on energy SNR per frame. And also, for the MSS_WPT [1] and the NMSS_WPT [1], we use the noise estimation method by first-order in [4], with the forgetting factor 0.9 and threshold 2.5, and the spectrum magnitude averaging method using the pause region detected by VAD [9].

Figure 2 and figure 3 show the average of improved segmental SNR and the average of WSS, observed from the speech enhanced by the proposed method and the comparison ones. In the average of improved segmental SNR, we observe that the performance of PM (12.20 dB) is better than that of MSS_WPT (5.00 dB) and NMSS_WPT (8.33 dB), respectively. As shown in figure 3, the MSS_WPT yield smaller value than the PM in the case of white Gaussian noise. In WSS, the smaller value means the better performance. However, it is hard to say that the MSS_WPT is better than the PM, because the difference between WSS values of the MSS_WPT and the PM is small and the MSS_WPT shows the better performance than the PM, only in the case of white Gaussian noise. Figure 4 and figure 5 are the spectrograms of enhanced speech by MSS_WPT, NMSS_WPT, and PM. As shown in both figures, we can observe that the musical tone in the enhanced speech by PM remains less compared with that by the each comparison method. For the subjective test, we use the PSS and the MOS. Table 1 shows that the PM gets higher score than any other methods.

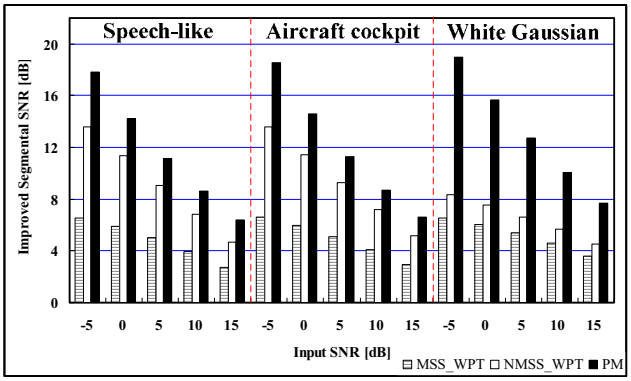

Figure 2: Comparison of improved segmental SNR

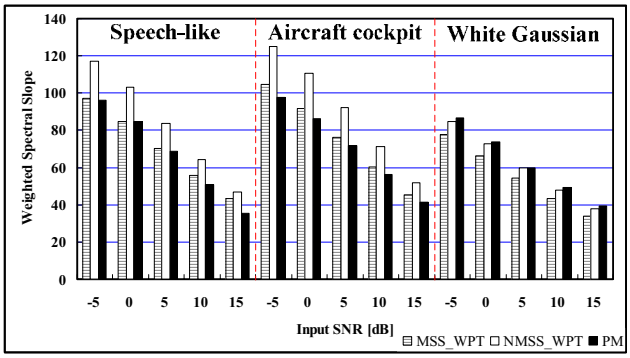

Figure 3: Comparison of weighted spectral slope

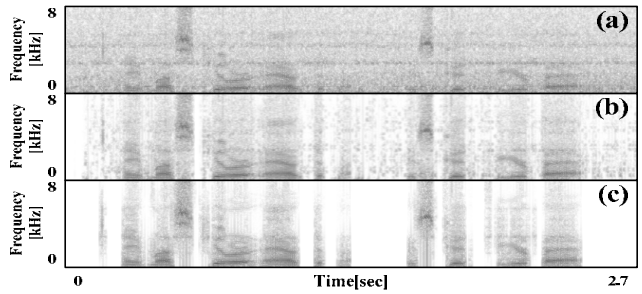

Figure 4: Spectrograms of enhanced speech corrupted by aircraft cockpit noise at SNR 5 dB. (a) MSS_WPT, (b) NMSS_WPT, (c) proposed method

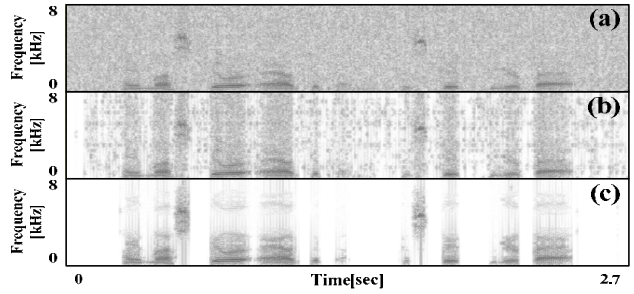

Figure 5: Spectrograms of enhanced speech corrupted by white Gaussian noise at SNR 0 dB. (a) MSS_WPT, (b) NMSS_WPT, (c) proposed method

Table 1: MOS for various enhancement methods

enhancement method	MOS
MSS_WPT	1.16
NMSS_WPT	1.92
PM	**2.84**

4. CONCLUSION

In this paper, we proposed the enhancement method based on WPT and BFRL, which has the following advantages.

1) Accurate estimation of parameters for speech enhancement using BFRL
2) Adaptive noise estimation using noise information of previous frame only, without the statistical information of the pause region
3) Efficient speech enhancement using the differential adaptive oversubtraction factor

5. REFERENCES

[1] B. Carnero and A. Deygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithm," *IEEE Trans. Signal Processing*, vol. 47, pp. 1622-1635, Jun. 1999.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustic Speech Signal Processing*, vol. ASSP-27, pp. 113-120, Apr. 1979.

[3] B. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE ICASSP*, pp. 208-211, Apr. 1979.

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

[5] S. Chang, S. Jung, Y. Kwon, and S. Yang, "Speech enhancement using wavelet packet transform," *IEEE ICSLP 2002,* pp. 1809-1812, 2002.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE ASSP*, vol. 33. pp. 443-445, Apr. 1985.

[7] S. Burrus, R. Gopinath, and H. Guo, *Introduction to wavelets and wavelet transforms*, Upper Saddle River, NJ: Prentice-Hall, 1998.

[8] T. Moon and W. Stirling, *Mathematical methods and algorithms for signal processing*, Upper Saddle River, NJ: Prentice-Hall, 2000.

[9] L. Lamel, L. R. Rabiner, A. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE ASSP*, vol. 29, pp. 777-785, 1981.

[10] J. Deller, J. Proakis, and J. Hansen, *Discrete-time processing of speech signals*, Englewood Cliffs, NJ: Prentice-Hall, 1993.