A Novel Speech Enhancement Method Using Power Spectra Smooth in Wiener Filtering

Feng Bao, Hui-jing Dou, Mao-shen Jia and Chang-chun Bao
Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China, 100124
E-mail: baofeng@emails.bjut.edu.cn, dhuijing@bjut.edu.cn, jiamaoshen@bjut.edu.cn, baochch@bjut.edu.cn

Abstract— **In this paper, we propose a novel speech enhancement method by using power spectra smooth of the speech and noise in Wiener filtering based on the fact that a priori SNR in standard Wiener filtering reflects the power ratio of speech and noise in frequency bins. This power ratio also could be approximated by the smoothed spectra of speech and noise. We estimate the power spectra of noise and speech by means of minima controlled recursive averaging method and spectral-subtractive principle, respectively. Then, the linear prediction analysis is used to smooth power spectra of the speech and noise in frequency domain. Finally, we utilize cross-correlation between the power spectra of the noisy speech and noise to modify gains of the power spectra for further reducing noise in silence and unvoiced segments. The objective test results show that the performance of the proposed method outperforms conventional Wiener Filtering and Codebook-based methods.**

Index Terms—Speech enhancement, Wiener filtering, Linear prediction

I. Introduction

The goal of speech enhancement is to remove a certain amount of noise from a noisy speech signal while keeping the speech component and reducing speech distortion as much as possible. Although the research on speech enhancement has been developed for a long time, the issues such as distortions to the original speech and residual noise sometimes created by the enhancement algorithms remain unsolved.

Currently, single channel speech enhancement, such as Wiener filtering method in a frequency domain [1], spectral-subtractive algorithm [2] and statistical-model-based method [3] [4], have become the core algorithms or baseline algorithms for further research. These baseline algorithms show the good performance for stationary noise. However, they may cause more artificial noise for non-stationary noise, especially in silence or unvoiced segments. After that, an alternative method based on a priori information about spectral shapes of speech and noise to substitute for iterative algorithm is codebook driven Wiener filtering (CDWF) [5]. Since the gains derived from shape codebook of spectra with respect to speech and noise could not match clean speech and real noise very well, it retains a lot of fluctuant background noise, especially in silence and unvoiced segments. Although this Wiener method and statistical-model-based method are effective, its performance entirely depends on the posterior SNR. Once the posterior SNR is not accurate for non-stationary noise, the quality of the enhanced speech will be degraded. In order to compare the performance concerned later, we name this method as the baseline Wiener filtering (BWF).

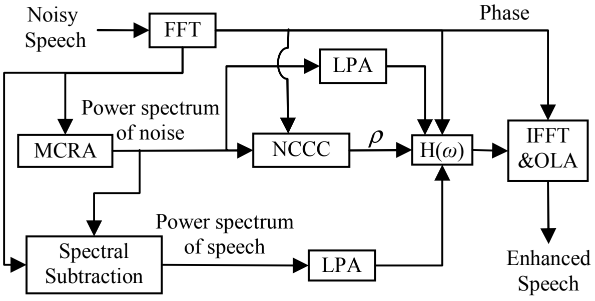

Fig 1. Block Diagram of the proposed method

In this paper, based on a frequency domain Wiener filtering principle [1], we only use the smoothed power spectra of speech and noise to construct Wiener filter instead of using a priori SNR estimation used in [3] and priori information about the spectra of speech and noise used in [5]. We name this method as the power spectrum smooth based Wiener filtering (PSSBWF) in this paper for convenience.

The main structure of the proposed method is shown in Fig.1. First, by taking the Fast Fourier Transform (FFT) of the input signal, we can get the magnitude and phase spectra of noisy speech. Then, minima controlled recursive averaging (MCRA) [6] method is applied to estimate the power spectrum of noise. By combining the power spectrum of noisy speech, the power spectrum of speech is estimated by spectral subtractive principle [2]. Considering the overestimation or underestimation of spectral subtractive algorithm, next, we need to smooth power spectra of speech and noise by linear prediction analysis (LPA) [7] for obtaining spectral shapes and gains of the speech and noise, respectively. After that, the Wiener Filter in a frequency domain is constructed by spectral shapes and gains of the speech and noise. In order to suppress the noise in silence or unvoiced segments, the gains of the power spectra about noise and speech used in the constructed Wiener Filter is modified frame by frame with the normalized cross-correlation coefficients (NCCC) between the power spectra of noisy speech and noise [8]. Finally, the enhanced speech is obtained by taking an inverse FFT and an overlap-and-add (OLA).

978-616-361-823-8 © 2014 APSIPA

The remainder of this paper is organized as follows. In Section Ⅱ, the baseline Wiener filtering method is briefly described. Wiener filter reconstruction based on the smoothed spectra of speech and noise is given in Section Ⅲ. The performance evaluation is presented in Section Ⅳ. Finally, we give the conclusions in Section Ⅴ.

II. Baseline Wiener Filtering

The standard Wiener filter in the frequency domain is generally expressed as follows:

$$H(\omega) = P_s(\omega) \big/ \left(P_s(\omega) + P_n(\omega)\right) \tag{1}$$

where $P_s(\omega)$ and $P_n(\omega)$ are the power spectra of speech and noise, respectively.

Let $\xi_k = P_s(\omega)/P_n(\omega)$, the equation (1) can be written as

$$H(\omega) = \xi_k \big/ (\xi_k + 1) \tag{2}$$

where ξ_k is defined as a priori SNR. Thus, the solution of Wiener filter depends on ξ_k, that is, it depends on the power spectra of speech and noise, which we do not have beforehand. Fortunately, Y. Ephraim and D. Malah gave an effective method for estimating ξ_k in [3]. In order to compare the performance with the proposed method, we name this kind of Wiener filtering with a priori SNR estimation as the baseline Wiener filtering (BWF) for comparison in this paper.

III. Power Spectrum Smooth Based Wiener Filtering

A. Power Spectra Smooth of the Noise and Speech

From equation (2), we can see that a priori SNR ξ_k reflects the power ratio of speech and noise in frequency bins. This indicates that such kind of power ratio also can be approximated by the smoothed spectra of speech and noise.

Based on LPA, the smoothed spectrum of speech or noise can be represented by

$$A(\omega) = G \Bigg/ \left(1 - \sum_{i=1}^{p} a_i e^{-j\omega i}\right) \tag{3}$$

where G is gain, and $\{a_i\}_{i=1,2,\cdots,p}$ are the prediction coefficients that represent the shape of power spectrum, p is predictor order. In the frequency domain, the smoothed spectrum can be regarded as a fitting process of power spectrum. In this paper, p is set to 10 for speech and noise.

Generally, the parameters G and $\{a_i\}$ are estimated in time domain by LPA. But we do not know the speech and noise signals in time domain in advance. So, in this paper, we try to use the estimated power spectrum $P(\omega)$ of speech or noise instead of time-domain signal to estimate the auto-correlation coefficients $r(i)$ used in LPA. According to the Wiener-Khintchine theorem, $r(i)$ can be expressed as follows:

$$r(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) \cos(i\omega) d\omega, \quad i = 1, 2, .., p \tag{4}$$

Since $P(\omega)$ is a real even function, the equation (4) only has cosine terms. By using Levinson-Durbin recursive algorithm [7], we can obtain the prediction coefficients $\{a_i\}$, and the residual energy of $P(\omega)$ can be calculated as follows:

$$E_p = r(0) - \sum_{i=1}^{p} \alpha_i r(i) \tag{5}$$

The gain of speech or noise can be obtained as follows:

$$G = \sqrt{E_p} \tag{6}$$

The key problem for solving equation (3) in frequency domain is how to obtain the power spectra estimation of noise and speech. In this paper, we adopted a direct method, that is, the power spectrum of noise $P_n(\omega)$ is obtained by effective MCRA method [6], and the power spectrum of speech $P_s(\omega)$ is obtained by subtracting $P_n(\omega)$ from the power spectrum of noisy speech. An main reason why we use the smoothed spectra instead of using the estimated $P_s(\omega)$ and $P_n(\omega)$ is that $P_s(\omega)$ estimated by spectral-subtractive principle often causes overestimation or underestimation. This will lead to a serious distortion for the enhanced speech.

In order to avoid a negative estimation of $P_s(\omega)$ owing to inaccuracies in estimating the noise spectrum, we calculate $P_s(\omega)$ as follows:

$$P_s(\omega) = \left(|Y(\omega)| - |N(\omega)|\right)^2 \tag{7}$$

where $|Y(\omega)|$ and $|N(\omega)|$ are the magnitude spectra of the noisy speech and the estimated noise, respectively.

B. Construction of Wiener Filter

Based on the smoothed spectra of speech and noise, we can construct the following Wiener filter:

$$\bar{H}(\omega) = \bar{P}_s(\omega) \big/ \bar{P}_s(\omega) + \bar{P}_n(\omega) \tag{8}$$

where $\bar{P}_s(\omega)$ and $\bar{P}_n(\omega)$ are the smoothed power spectra of speech and noise, respectively. They are represented as follows, respectively:

$$\bar{P}_s(\omega) = G_s^2 \Bigg/ \left(\left|1 - \sum_{i=1}^{p} a_i^{(s)} e^{-j\omega i}\right|^2\right), \; \bar{P}_n(\omega) = G_n^2 \Bigg/ \left(\left|1 - \sum_{i=1}^{p} a_i^{(n)} e^{-j\omega i}\right|^2\right) \tag{9}$$

where G_s and G_n are the gains of the power spectra for noise and speech, respectively. $\left\{a_i^{(s)}\right\}_{i=1,2,\cdots,p}$ and $\left\{a_i^{(n)}\right\}_{i=1,2,\cdots,p}$ are the prediction coefficients of the power spectra for speech and noise, respectively. When noisy speech passes through this filter, the enhanced speech sounds soft and natural for listening. We cannot hear the vacuum feeling and fluctuant noise produced by BWF and CDWF aforementioned. But we can feel that the noise level is a little larger in silence and unvoiced segments. In order to further reduce the noise in these segments, we exploit the correlation between the power spectra of noisy speech and the estimated noise to modify gains G_s and G_n used in equation (9). The modified gains are given as follows, respectively:

$$\overleftrightarrow{G}_n^2 = \rho \cdot G_n^2, \quad \overleftrightarrow{G}_s^2 = (1 - \rho) \cdot G_s^2 \tag{10}$$

where ρ is the normalized cross-correlation coefficient (NCCC) determined by

$$\rho = \sum_{k=0}^{M-1} |Y(k)||N(k)| \Bigg/ \sqrt{\sum_{k=0}^{M-1} |Y(k)|^2 \sum_{k=0}^{M-1} |N(k)|^2} \tag{11}$$

where $|Y(k)|$ is magnitude spectrum of the noisy speech, $|N(k)|$ is magnitude spectrum of the noise. M is FFT size, k is the index of frequency bins.

The calculation of ρ is based on the fact that there is correlation to some extent between the spectra of the noisy speech and noise in unvoiced and silence segments. Here, ρ varies frame by frame.

Fig. 2 shows the variation of ρ versus clean speech waveform. In the ideal case (the estimated noise nearly matches the noisy speech), the value of ρ will close to 1 in silence or unvoiced segments. While, due to the deviation of noise estimation, the ρ approaches a larger value that is less than 1 in silence or unvoiced segments. For example, in Fig. 2, the ρ is about 0.9 in silence or unvoiced segments. It still can be used for modifying the gain of noise signal to suppress the background noise. On the contrary, ρ decrease to a small value in voiced segments. Thus, it can be utilized to adjust the gain of speech signal in order to increase the energy of speech. The noise existed in silence segments and in unvoiced segments is almost suppressed and no artifacts are produced in the voiced segments by this kind of modification.

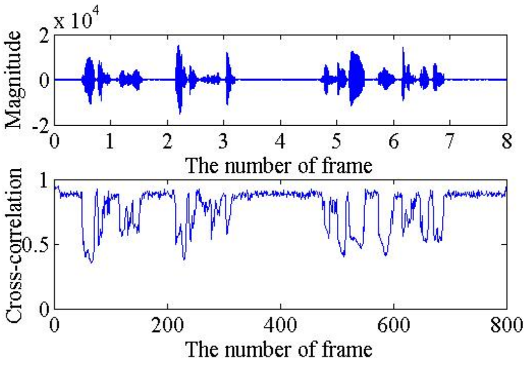

Fig. 2. The variation of ρ vs. clean speech waveform

Fig. 3 shows an example for short-time energy comparison at 10 dB SNR for street noise. We found that the PSSBWF with NCCC has the lowest noise level in silence and unvoiced segments. This makes us more comfortable for listening compared with BWF and CDWF methods. Particularly, the NCCC plays an important role in reducing noise level.

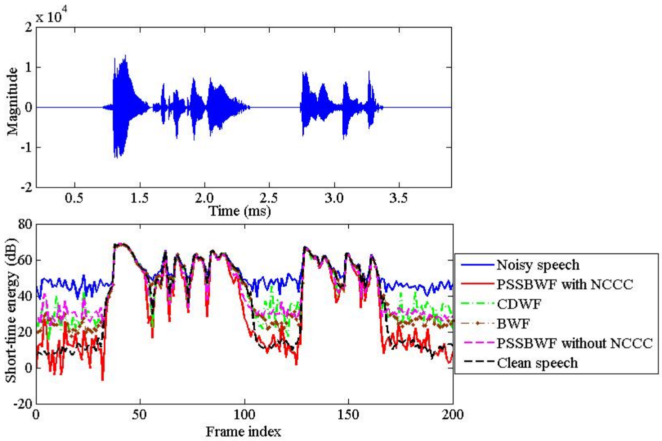

Fig. 3. Short-time energy comparison of the enhanced speech between BWF, CDWF and PSSBWF

IV. Performance Evaluation

In performance evaluation, the test speech is selected from NTT database including nine utterances from four female speakers and five male speakers. The length of each utterance is 8s. Four types of noise including white, babble, office and street are chosen from NOISE 92 database. The input signal to noise ratio (SNR) is defined as 0dB, 5dB and 10dB, respectively. The sampling rate of speech signal is 8 kHz. We used average segmental signal-to-noise ratio (SSNR) [9], average log-spectral distortion (LSD) [10] and Perceptual Evaluation of Speech Quality (PESQ) [11] to do the objective measurement of the enhanced speech quality. The SSNR is defined as follows [9]:

$$SSNR = \frac{1}{M}\sum_{j=1}^{M} 10\log_{10}\left(\sum_{n=1}^{N} x^2(n) \Big/ \sum_{n=1}^{N}[x(n)-\hat{x}(n)]^2\right) \tag{12}$$

where M is the number of frame used for test. N is the length of frame. Here, the length of frame is 160 samples (20ms). X(n) denotes clean speech signal and $\hat{X}$ indicates the enhanced speech or noisy speech. The LSD is given by [10]:

$$d_{LSD} = \frac{1}{M}\sum_{l=0}^{M-1}\sqrt{\frac{1}{N_{fft-size/2}}\sum_{k=0}^{N_{fft-size/2}-1}\left[10\log_{10}\left(\left|\hat{X}(l,k)\right|^2 \Big/ \left|X(l,k)\right|^2\right)\right]^2} \tag{13}$$

where l is the frame index, k is the index of frequency bins, M is the number of frame, $N_{fft\text{-}size}$ *is* the FFT size. Here, $N_{fft-size} = 512$. $|X(l,k)|^2$ and $|\hat{X}(l,k)|^2$ denote the power spectra of the clean speech and the enhanced speech, respectively.

Table 1 Test Result of SSNR Improvement

Noise Type	Method	10dB	5dB	0dB
white	Noisy	-	-	-
	BWF	10.06	14.09	15.41
	CDWF	6.67	7.18	7.64
	Proposed	14.29	16.41	17.78
babble	Noisy	-	-	-
	BWF	6.75	8.40	9.01
	CDWF	4.13	4.81	5.37
	Proposed	11.18	12.73	14.23
Street	Noisy	-	-	-
	BWF	9.98	12.15	12.67
	CDWF	8.77	9.69	11.05
	Proposed	13.78	16.23	18.25
Office	Noisy	-	-	-
	BWF	9.93	10.60	11.18
	CDWF	5.78	6.86	7.46
	Proposed	12.67	14.59	14.21

Table 1 shows the Test Result of SSNR Improvement compared with BWF and CDWF methods. The same MCRA noise estimation method [6] is used for BWF method and a priori SNR in equation (2) is obtained by [3] for BWF method. For the CDWF method, the size of shape codebook in terms of 10-dimension line spectrum frequencies (LSF) about speech spectrum is 1024 (10bit), the size of shape codebooks in terms of 10-dimension LSF about the spectra of white noise, street noise, office noise and babble noise are 8 (3bit), 8 (3bit), 8 (3bit) and 16 (4bit), respectively. These priori shape codebooks of the spectra are trained by generalized Lloyd algorithm (GLA) [12] based on weighting distortion measure [13]. From Table 1, we can find that the proposed method produces a higher average SSNR improvement than other two methods for three kinds of input SNRs. The lower the SNR, the higher is its average SSNR improvement. From Table 2, we also see that the LSD of the proposed method is the lowest for four types of noise under three input SNR conditions.

Table 2 Test Result of LSD

Noise Type	Method	10dB	5dB	0dB
white	Noisy	14.09	16.14	18.31
	BWF	7.09	8.30	9.57
	CDWF	10.46	12.03	13.78
	Proposed	6.96	8.03	9.26
babble	Noisy	10.60	12.35	14.24
	BWF	6.79	8.29	9.90
	CDWF	8.82	10.30	11.92
	Proposed	6.24	7.49	8.95
Street	Noisy	9.26	10.91	12.68
	BWF	5.11	6.33	7.76
	CDWF	7.00	8.31	9.72
	Proposed	5.00	5.83	7.05
Office	Noisy	9.56	11.24	13.06
	BWF	6.04	7.36	8.85
	CDWF	7.79	9.20	10.74
	Proposed	5.45	6.67	7.80

PESQ test result listed in Table 3 shows that three kinds of methods all improve the speech quality in a certain degree. Although the PESQ result of proposed method is a little worse than that of BWF method, the auditory perception of the proposed method looks much soft and natural than BWF which has vacuum feeling. The reason why causes this phenomenon is that the proposed method slightly suppresses some speech components which have weak energy in some transition segments. However, the proposed method is almost better than CDWF for the PESQ test with three different kinds of SNRs.

Table 3 Test Result of PESQ

Noise Type	Method	10dB	5dB	0dB
white	Noisy	1.97	1.60	1.36
	BWF	2.67	2.32	2.04
	CDWF	2.48	2.14	1.72
	Proposed	2.51	2.22	1.85
babble	Noisy	2.50	2.18	1.79
	BWF	2.69	2.36	1.91
	CDWF	2.46	2.13	1.72
	Proposed	2.66	2.34	1.99
Street	Noisy	2.95	2.65	2.30
	BWF	3.22	2.97	2.70
	CDWF	3.09	2.85	2.58
	Proposed	3.07	2.89	2.55
Office	Noisy	2.76	2.41	2.02
	BWF	2.90	2.61	2.23
	CDWF	2.84	2.53	2.15
	Proposed	2.84	2.55	2.18

V. CONCLUSIONS

In this paper, a novel speech enhancement method based on power spectra smooth of speech and noise is proposed. The smoothed spectra of speech and noise can effectively describe the power ratio of speech or noise in frequency bins, and makes the Wiener filtering easier for speech enhancement. Also, the introduction of the NCCC makes the constructed Wiener filter more effective for removing background noise existed in silence or in unvoiced segments. The objective test results show that the proposed method has a better performance, comparing with the reference methods, especially for reducing the vacuum feeling and fluctuant background noise.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No. 61471014, No. 61201197), and the Scientific Research Project of Beijing Educational Committee (Grant No. KM201310005008), the Ph.D. Programs Foundation of Ministry of Education of China (Grant No. 20121103120017).

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, Enhancement and bandwidth compression of noisy speech, Proc. IEE, vol. 67, no. 12, pp.1586-1604, Dec.1979.

[2] Boll, S.F, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process., 27(2),113-120. 1979

[3] Y. Ephraim and D. Malah, Speech Enhancement Using a Minimum Mean Square Error Short-time Spectral Amplitude Estimator, IEEE Tran. Acoust. Speech Signal Process., 32(6), 1109-1121. 1984,

[4] Loizou, P. Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Speech Magnitude Spectrum, IEEE Trans. Speech Audio Process., 13(5), 857-869, 2005.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, Codebook driven short-term predictor parameter estimation for speech enhancement, IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 1, pp. 163–176, Jan. 2006.

[6] Cohen, I. Noise estimation by minima controlled recursive averaging for robust speech enhancement, IEEE Signal Process. Lett., 9(1), 12-15. 2002.

[7] Lawrence R. Rabiner and Ronald W. Schafer. Theory and Applications of Digital Speech Processing, Pearson Education, Inc., 2011

[8] Feng Bao, Hui-jing Dou, Mao-shen Jia and Chang-chun Bao, Speech Enhancement Based on a Few Shapes of Speech Spectrum, ChinaSIP 2014, pp. 90-94.

[9] Quackenbush, S. R., Barnwell, T. P., Clements, M. A., Objective Measures of Speech Quality. Englewood Cliffs, NJ: Prentice Hall, 1988.

[10] Abramson, A., Cohen, I., Simultaneous Detection and Estimation Approach for Speech Enhancement. IEEE Trans. Speech Audio Process., 15(8), 2348-2359, 2007.

[11] ITU-T, Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders, 2001.

[12] Y.Linde, A. Buzo, and R. M .Gray, “An algorithm for vector quantizer design,” IEEE Trans. Commun, vol. COM-28, no. 1, pp 84-95, Jan. 1980.

[13] K. K. Paliwal and B. S. Atal. Efficient Vector Quantization of LPC Parameters at 24 bits/frame. IEEE Transactions on Speech and Audio Processing, 1993, 1(1), 3-14.