Speech Enhancement Using Optimized Wiener Filter for VoIP Speech Codecs

Seungho Han[1], Sangbae Jeong[1], Heesik Yang[1], Jinsul Kim[2], Won Ryu[2], and Minsoo Hahn[1]

[1] Speech and Audio Information Laboratory, Information and Communications University, 103-6, Munji-dong, Yuseong-gu, Daejeon, 305-732, Republic of Korea
{space0128, sangbae, sheik, mshahn}@icu.ac.kr
[2] Convergence Network Inter-working Technology Team, BcN Service Research Group, Broadband Convergence Network Research Division, Electronics and Telecommunications Research Institute, 161, Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Republic of Korea
{jsetri, wlyu}@etri.re.kr

Abstract

A speech enhancement technique is indispensable to achieve acceptable speech quality in VoIP systems. This paper proposes a Wiener filter optimized to the estimated SNR of noisy speech for speech enhancement. The proposed noise reduction method is applied as pre-processing before speech coding. The performance of the proposed method is evaluated by the PESQ in various noisy conditions. In this paper, G.711, G.723.1, and G.729A VoIP speech codecs are used for the performance evaluation. The PESQ results show that the performance of our proposed noise reduction scheme outperforms those of the noise suppression in the IS-127 EVRC and the noise reduction in the ETSI standard for the advanced distributed speech recognition front-end.

Keywords: speech enhancement, noise reduction, speech codec, VoIP

1. Introduction

Speech quality is significantly degraded in the presence of background noise. A speech enhancement or noise reduction technique therefore is essential to suppress the background noise. Various methods such as spectral subtraction schemes, and adaptive filtering techniques including Wiener and Kalman filtering have been proposed for noise reduction [Boll (1979)], [Ephraim et. Al. (1984)], [Gannot et. Al. (1998)], [Hayes (1996)], [Martin (2001)], [McAulay et. Al. (1980)], [Paliwal et. Al. (1987)], [Widrow et. Al. (1985)]. The noise reduction techniques are used to a wide range of applications such as communication, automatic speech recognition, and sound source localization systems.

VoIP services have been rapidly developed because of cost savings and the attractive possibility of growth. Background noise is one of the significant factors degrading the QoS in the VoIP. Because voice signals are usually acquired by one microphone in VoIP systems, a single microphone-based noise reduction technique is required for the systems. Since the delay is also the important QoS factor, the noise reduction should not induce intolerable delay. In this paper, we choose a Wiener filter-based noise reduction scheme because it is one of the popular approaches and shows a basically high performance. This paper proposes the Wiener filter optimized to the estimated SNR of speech for VoIP speech codecs. The performance is objectively evaluated by the ITU-T P.862 PESQ (Perceptual Evaluation of Speech Quality) [ITU-T Rec. P.862 (2001)]. Various noisy conditions including white Gaussian, office, babble, and car noises are considered. The performance of the proposed method is compared with those of the noise reduction methods in the IS-127 EVRC (Enhanced Variable Rate Codec) [EVRC (1996)] and in the ETSI (European Telecommunications Standards Institute) standard for the distributed speech recognition front-end [ETSI ES 202 212 (2005)]. The noise reduction block is embedded in the EVRC while it is not in the VoIP speech codecs. The ETSI noise reduction based on the 2 stage Wiener filter generates 40 msec algorithmic delay though it shows a high performance. And, the noise reduction is applied as pre-processing of VoIP speech codecs such as G.711, G.723.1, and G.729A.

The paper is organized as follows. Section 2 introduces the derivation of the Wiener filtering algorithm. Section 3 proposes our Wiener filter-based speech enhancement. Some experiments are described and the performance evaluation results are shown in section 4. Finally, section 5 draws the conclusions including further studies.

2. Derivation of Wiener Filtering Algorithm

For a noncausal IIR (Infinite Impulse Response) Wiener filter, a clean speech signal *d(n)*, a background noise *v(n)*, and an observed signal *x(n)* can be expressed as

$$x(n) = d(n) + v(n)\,. \tag{1}$$

It is assumed that *d(n)* and *v(n)* are jointly wide-sense stationary and uncorrelated. A Wiener filter is designed to minimize the mean square error

$$\xi = E\left[\left(d(n) - \hat{d}(n)\right)^2\right] = E\left[e^2(n)\right] \tag{2}$$

where $\hat{d}(n)$ is the output of the Wiener filter expressed as

$$\hat{d}(n) = \sum_{l=-\infty}^{\infty} w(l)x(n-l)\,. \tag{3}$$

Finally, the frequency response of the Wiener filter [Hayes (1996)] becomes

$$W(e^{jw}) = \frac{P_d(e^{jw})}{P_v(e^{jw}) + P_d(e^{jw})} = \frac{\zeta(e^{jw})}{1+\zeta(e^{jw})} \quad \textbf{(4)}$$

in which $\zeta(e^{j\omega})$ is considered as the SNR (Signal-to-Noise Ratio) defined by

$$\zeta(e^{jw}) = \frac{P_d(e^{jw})}{P_v(e^{jw})}. \quad \textbf{(5)}$$

3. Proposed Wiener Filter

Since a noncausal IIR filter is unrealizable in practice, we propose a causal FIR (Finite Impulse Response) Wiener filter. Figure 1 shows the proposed noise reduction process.

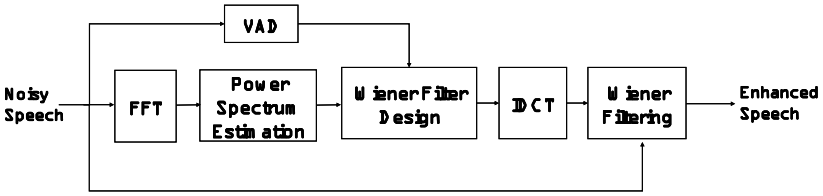

Figure 1. *Noise reduction process*

The speech enhancement is processed frame-by-frame. The processing frame having 80 samples is the current input frame. Total 100 samples, i.e., the current 80 and the past 20 samples, are used to compute the power spectrum of the processing frame. In the first frame, the past samples are initialized to zero. For the power spectrum analysis, the signal is windowed by the 100 sample-length asymmetric window $h(n)$ whose center is located at the 70th sample as follows.

$$h(n) = \begin{cases} 0.5 - 0.5\cos\left(\dfrac{2\pi n}{139}\right) & ;n < 70 \\ \cos\left(\dfrac{2\pi(n-70)}{119}\right) & ;else \end{cases} \quad \textbf{(6)}$$

The length and the center of the asymmetric window are empirically chosen to make the algorithm produce the best performance. The signal power spectrum is computed for this windowed signal using 256-FFT. In the Wiener filter design, the noise power spectrum is updated only for non-speech intervals by the decision of VAD (Voice Activity Detection) while the previous noise power spectrum is reused for speech intervals. And the speech power spectrum is estimated by the difference between the noise power and the signal power spectrum.

With these estimated power spectra, the proposed Wiener filter is designed. In our proposed Wiener filter, the frequency response is expressed as

$$W(k) = \frac{\zeta^{\alpha}(k)}{1+\zeta^{\alpha}(k)}, 0 < \alpha \leq 1 \quad \textbf{(7)}$$

and $\zeta(k)$ is defined by

$$\zeta(k) = \frac{P_d(k)}{P_v(k)} \quad \textbf{(8)}$$

where k is the frequency bin, $\zeta(k)$, $P_d(k)$, and $P_v(k)$ are the SNR, the speech power spectrum, and the noise power spectrum, respectively. Therefore, filtering is controlled by the parameter α. For $\zeta(k)$ greater than one, as α is increased, $\zeta^{\alpha}(k)$ is also increased while $\zeta^{\alpha}(k)$ is decreased for $\zeta(k)$ less than one. The signal is more strongly filtered out to reduce the noise for smaller $\zeta^{\alpha}(k)$. On the other hand, the signal is more weakly filtered with little attenuation for larger $\zeta^{\alpha}(k)$.

To analysis the effect of α, we evaluate the performances for α values from 0.1 to 1. The performance is evaluated not for the coded speech but for the original speech in white Gaussian conditions. The other experimental surroundings are explained in section 4. As α is increased up to 0.7, the performance is improved. The performance becomes worse after that. Table 1 shows the PESQ scores according to the SNR when α is set to 0.6 and 0.7, from now on, will be referred as case A and case B, respectively. The performance of the case A is better for high SNR but worse for low SNR. Thus, we can adaptively select the optimal α according to the estimated SNR by a data-driven approach.

Table 1. *Comparison of PESQ scores*

SNR(dB)	20	15	10	5	0
Case A	3.25	2.94	2.61	2.23	1.76
Case B	3.24	2.90	2.60	2.26	1.81
Proposed	3.26	2.95	2.63	2.26	1.78

The codebook is trained for deciding the optimal α to the estimated SNR. First, the estimated SNR mean is calculated for the current frame. Second, the spectral distortion is measured with the log spectral Euclidean distance *D* defined as

$$D = \frac{1}{L}\sqrt{\sum_{k=0}^{L-1}\left[\log\left|X_{ref}(k)\right| - \log\left\{\left|X_{in}(k)\right|W(k)\right\}\right]^2} \quad \textbf{(9)}$$

where k is the index of the spectral bins, L is the total number of the spectral bins, $|X_{ref}(k)|$ is the spectrum of the clean reference signal, and $|X_{in}(k)|W(k)$ is the noise-reduced signal spectrum after filtering with the designed Wiener filter. Third, for the frame, optimal α is searched to minimize the distortion. For all frames, the estimated SNR means of all bins with the optimal α are clustered by the LBG algorithm. Finally, the optimal α for each cluster is decided by averaging all α values in the cluster. The clustered SNR means and the corresponding optimal α consist of the SNR codebook of which size is 64.

When the Wiener filter is designed, the optimal α is searched by comparing the estimated SNR mean of all bins with the codeword of the cluster as shown in Figure 2.

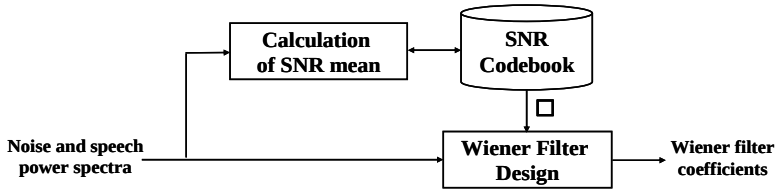

***Figure 2**. Wiener filter design by searching for the optimal α*

The designed Wiener filter coefficients in the frequency domain are transformed into the time-domain ones by the IDCT (Inverse Discrete Cosine Transform). Finally, the noise is reduced by the convolution sum between the impulse response of the proposed Wiener filter and the noisy speech. Because the proposed Wiener filter is a causal filter, the algorithmic delay is unavoidable after the convolution. However, the delay may be approximately the half of the filter length and almost negligible compared to the total VoIP delay.

When our proposed algorithm is applied, the highest performance is shown for all conditions except in 0 dB, an inconsiderable condition in practice, according to SNR in Table 1. Thus, the proposed Wiener filter effectively reduces the noises by optimizing to the estimated SNR.

4. Experiments and Results

The objective speech quality is evaluated by the PESQ to confirm the performance. After comparing an original signal with its degraded signal, the PESQ gives us reliable estimation of the subjective measurement as an MOS-like value from -0.5 to 4.5.

One hundred mono spoken sentences with 16 bit resolution and 8 kHz sampling rate are used as the clean speech in our experiments. The utterances are spoken by 2 males and 2 females. The duration of each utterance is about 10 seconds. As the open test,

40 spoken sentences are used to train the SNR codebook and the others, to evaluate the performance. In the training process, the parameter α is varied from 0.6 to 0.7. The number of the proposed Wiener filter coefficients is 65. Therefore, the algorithmic delay is 4 msec.

Zero mean white Gaussian noises are generated and office, babble, and car noises are recorded with a Sony digital audio tape recorder. Office noises are collected in the room where many computers and an air-conditioner are operated. Babble noises are collected in a cafeteria where many people make a noise. Car noises are collected in a driving car at 70 km/h speed with the windows closed but the air-conditioner on.

The noise signals are added to the clean speech ones to produce noisy ones with the SNR of 0, 5, 10, 15, and 20 dB. Therefore, total 800 noisy spoken sentences are trained because there are 5 SNR levels, 40 speech utterances, and 4 types of noises. For each test codec, total 1200 noisy spoken sentences are tested. The noise is reduced as pre-processing before encoding the speech in a codec. Final decoded speech is evaluated by the PESQ. Popular VoIP speech codecs such as G.711, G.723.1, and G.729A are tested.

To verify the performance, our results are compared with those of the noise suppression in the IS-127 EVRC and the noise reduction in the ETSI standard for the advanced distributed speech recognition front-end. The ETSI noise canceller generates 40 msec buffering delay while there is no buffering delay in the EVRC noise canceller.

Table 2, 3, and 4 show the average PESQ results in G.711, G.723.1, and G.729A, respectively. In most noisy conditions, the proposed method yields higher PESQ scores than the others.

Table 2. *Average PESQ results in G.711*

	White Gaussian noise					Office noise				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
None	2.52	2.20	1.89	1.61	1.42	2.77	2.44	2.11	1.76	1.44
EVRC	3.07	2.73	2.32	1.83	1.39	3.13	2.78	2.42	2.02	1.64
ETSI	3.17	2.84	2.48	2.09	1.69	3.18	2.80	2.44	2.04	1.64
Proposed	**3.27**	**2.96**	**2.64**	**2.27**	**1.82**	**3.25**	**2.88**	**2.52**	**2.13**	**1.74**
	Babble noise					Car noise				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
None	2.73	2.40	2.11	1.79	1.51	3.61	3.28	2.97	2.64	2.28
EVRC	3.11	2.76	2.45	2.04	1.67	3.84	3.52	3.22	2.90	2.57
ETSI	3.14	2.76	2.48	2.11	1.73	3.85	3.60	3.26	2.92	2.55
Proposed	**3.18**	**2.81**	**2.49**	**2.13**	**1.75**	**3.89**	**3.65**	**3.30**	**3.01**	**2.63**

Table 3. *Average PESQ results in G.723.1 (6.3kbps)*

	White Gaussian noise					Office noise				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
None	2.64	2.34	2.03	1.70	1.44	2.87	2.58	2.28	1.91	1.54
EVRC	2.95	2.70	2.36	1.90	1.40	3.05	2.77	2.46	2.07	1.68
ETSI	3.07	2.85	2.55	2.20	1.78	3.09	2.82	2.51	2.12	1.72
Proposed	**3.10**	**2.85**	**2.59**	**2.27**	**1.86**	**3.12**	**2.84**	**2.52**	**2.15**	**1.78**
	Babble noise					Car noise				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
None	2.84	2.55	2.27	1.92	1.59	3.29	3.09	2.84	2.55	2.22
EVRC	3.04	2.77	2.48	2.11	1.72	3.40	3.21	2.97	2.68	2.36
ETSI	3.07	2.79	2.52	2.19	1.79	3.40	3.26	3.02	2.75	2.41
Proposed	**3.06**	**2.80**	**2.47**	**2.15**	**1.76**	**3.45**	**3.35**	**3.12**	**2.90**	**2.54**

Table 4. *Average PESQ results in G.729A*

	White Gaussian noise					Office noise				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
None	2.71	2.40	2.06	1.71	1.43	2.93	2.63	2.31	1.92	1.51
EVRC	2.98	2.74	2.40	1.93	1.40	3.09	2.80	2.48	2.08	1.67
ETSI	3.13	2.89	2.59	2.23	1.79	3.13	2.84	2.51	2.09	1.67
Proposed	**3.14**	**2.86**	**2.60**	**2.27**	**1.85**	**3.15**	**2.85**	**2.52**	**2.11**	**1.72**
	Babble noise					Car noise				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
None	2.90	2.60	2.30	1.94	1.59	3.36	3.15	2.87	2.57	2.23
EVRC	3.08	2.80	2.51	2.13	1.74	3.49	3.32	3.06	2.74	2.40
ETSI	3.11	2.83	2.54	2.19	1.77	3.46	3.30	3.03	2.73	2.37
Proposed	**3.10**	**2.83**	**2.47**	**2.14**	**1.75**	**3.52**	**3.40**	**3.15**	**2.89**	**2.51**

Table 5 shows the average PESQ gains for all types of noises. The superiority of our proposed method can be observed in all cases. When comparing with the EVRC noise suppression, the proposed method produces significant PESQ gains. Moreover, the proposed method is slightly better than the ETSI noise reduction which generates 40 msec buffering delay. Our proposed method is especially effective for the noise reduction of the SNR from 5 dB to 15 dB. The algorithm produces the largest gain in G.711 based on a waveform-based coding technique.

Table 6 shows the average PESQ gains for all test speech codecs. The proposed method shows the largest PESQ gain for the white Gaussian noises. For the babble noises, even though the PESQ gains by the proposed method are greater than those by the EVRC noise suppression, they become slightly less than those by the ETSI noise reduction.

Table 5. *Average PESQ gains for all types of noise*

	G.711				
SNR(dB)	20	15	10	5	0
EVRC	0.38	0.37	0.33	0.25	0.16
ETSI	0.43	0.42	0.40	0.35	0.24
Proposed	**0.49**	**0.50**	**0.47**	**0.44**	**0.32**
	G.723.1 (6.3kbps)				
SNR(dB)	20	15	10	5	0
EVRC	0.20	0.22	0.22	0.17	0.09
ETSI	0.25	0.29	0.30	0.29	0.23
Proposed	**0.27**	**0.31**	**0.32**	**0.34**	**0.28**
	G.729A				
SNR(dB)	20	15	10	5	0
EVRC	0.19	0.22	0.23	0.19	0.11
ETSI	0.23	0.27	0.28	0.28	0.21
Proposed	**0.25**	**0.29**	**0.30**	**0.32**	**0.27**

Table 6. *Average PESQ gains for all test VoIP speech codecs*

	White Gaussian noise					Office noise				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
EVRC	0.38	0.41	0.37	0.21	-0.03	0.23	0.23	0.22	0.19	0.17
ETSI	0.50	0.55	0.55	0.50	0.32	0.28	0.27	0.25	0.22	0.18
Proposed	**0.55**	**0.58**	**0.62**	**0.60**	**0.41**	**0.32**	**0.31**	**0.29**	**0.27**	**0.25**
	Babble noise					Car noise				
SNR(dB)	20	15	10	5	0	20	15	10	5	0
EVRC	0.25	0.26	0.25	0.21	0.15	0.16	0.18	0.19	0.19	0.20
ETSI	0.28	0.28	0.29	0.28	0.20	0.15	0.21	0.21	0.21	0.20
Proposed	**0.29**	**0.30**	**0.25**	**0.26**	**0.19**	**0.20**	**0.29**	**0.30**	**0.35**	**0.32**

5. Conclusions

This paper proposes a new Wiener filtering scheme optimized to the estimated noisy signal SNR to reduce additive noises. For the use in VoIP applications, the proposed algorithm produces a very short delay. The proposed speech enhancement is applied before encoding as pre-processing of VoIP speech codecs. The PESQ results show that the performance of the proposed approach is superior to the EVRC noise suppression and the ETSI noise reduction. Therefore, our proposed speech enhancement can be effectively used to reduce additive noises in VoIP applications.

Further studies may include the consideration of the other methods for speech distortion measure such as the Kullback-Liebler distance and the Bark spectral distance. And, the Wiener filter can be designed by the estimated SNR in each spectral bin while the proposed Wiener filter in this paper is designed by the estimated SNR mean of all spectral bins.

References

Boll S. F. (1979), *Suppression of acoustic noise in speech using spectral subtraction*, in IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-27, no. 2, pp. 113-120.

Enhanced Variable Rate Codec (EVRC). (1996), *Speech service option 3 for wide-band spectrum digital systems*.

Ephraim Y., Malah D. (1984), *Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator*, in IEEE trans. Acoust., Speech, Signal Process., vol. ASSP-32, no. 6, pp. 1109-1121.

ETSI ES 202 212. (2005), *Speech processing, transmission and quality aspects (STQ); distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithms; back-end speech reconstruction algorithm*.

Gannot S., Burshtein D., Weinstin E. (1998), *Iterative and sequential Kalman filter-based speech enhancement algorithms*, in IEEE Trans. Speech Audio Process., vol. 6, no. 4, pp. 373-385.

Hayes M. H. (1996), *Statistical Digital Signal Processing and Modeling*, John Wiley&Sons, Printed in the USA, ISBN 0-471-59431-8.

ITU-T Rec. P.862. (2001), *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codec*.

Martin R. (2001), *Noise power spectral density estimation based on optimal smoothing and minimum statistics*, in IEEE Trans. Speech Audio Processing, vol. 9, no. 5, pp. 504-512.

McAulay R. J., Malpass M. L. (1980), *Speech enhancement using a soft decision noise suppression filter*, in IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-28, no. 2, pp. 137-145.

Paliwal K. K., Basu A. (1987), *A speech enhancement method based on Kalman filtering*, in Proc. ICASSP-87: International Conference on Acoustics, Speech, and Signal Processing, Dallas, USA, pp. 177-180.

Widrow B., Stearns S. D. (1985), *Adaptive Signal Processing*, Englewood Cliffs, NJ 07632: Prentice Hall, printed in the USA, ISBN 0-13-004029-0.