A POST-PROCESSING TECHNIQUE FOR REGENERATION OF OVER-ATTENUATED SPEECH COMPONENTS

*Huijun Ding, Ing Yann Soon, Soo Ngee Koh, Chai Kiat Yeo**

School of Electrical and Electronic Engineering
*School of Computer Engineering
Nanyang Technological University, Singapore 639798
Email: {ding0032, eiysoon, esnkoh, asckyeo}@ntu.edu.sg

ABSTRACT

Despite the success of recent speech enhancement algorithms, the enhanced signals still suffer from undesirable speech distortion caused by over-attenuation of weak speech spectral components. In this paper, a post-processing technique based on the regeneration of both voiced and unvoiced speech is proposed to alleviate this problem. A non-linear transformation is first applied to a Wiener filtered speech and the transformed signal is multiplied by a pre-estimated spectral envelop to form the regenerated speech. The resulting speech is then obtained using a weighted combination of the regenerated speech components and the filtered speech. This process significantly improves the resulting speech quality as compared to the original filtered version. It results in speech that sounds less lowpassed. Also, the residual musical noise is significantly masked by the regenerated speech components. Objective measures show that the quality of the resulting speech is much closer to the clean speech as compared to the original Wiener filtered speech.

Index Terms— Speech enhancement, speech processing

1. INTRODUCTION

Single channel speech enhancement can be used either as a stand-alone system or a pre-processor for some ensuing high level tasks. It is an important research area that has been widely studied for many years. The main challenge of single channel speech enhancement arises from the insufficient information available to separate the underlying speech from the uncorrelated noise. In the past, many algorithms have been proposed to solve this problem, such as the spectral subtraction (SS) algorithm [1], the minimum-mean square error (MMSE) Estimator [2] and Wiener filter based algorithms [3, 4]. Many of these methods share a common principle, that is, using a product of the noisy speech and a spectral gain function $g(\omega)$ to recover the clean speech. The only difference among these algorithms is the use of different $g(\omega)$. Therefore the difference between the estimated clean speech and the original clean speech comprises two parts: speech distortion $[g(\omega) - 1]\, S(\omega)$, and noise distortion $g(\omega)N(\omega)$, where $S(\omega)$ and $N(\omega)$ are the spectra of the original clean speech and noise respectively. References [5, 6] discuss the tradeoff between speech distortion and noise reduction. The common problem is that the higher the degree of noise suppression, the higher is the amount of speech distortion. The suppression of speech components is especially noticeable for the unvoiced speech, where the weak speech components are removed together with the noise components.

Recently, a few algorithms have been proposed to reduce this phenomenon. Harmonic regeneration based approach presented in [7, 8] tries to obtain a suppression gain that varies according to the speech harmonics to achieve a good tradeoff between speech distortion and noise reduction. However, this algorithm only attempts to reduce distortion in the speech harmonics and it does nothing for the unvoiced speech. In [9], a post-processing method is proposed to be used for certain algorithms where only high frequency speech components are regenerated. In order to obtain a fully regenerated speech, we propose an algorithm to recover both the voiced and unvoiced speech in the entire frequency domain. In our technique, a smoothed envelop is first estimated to produce a continuous spectrum for regeneration. Then the excitation signal is calculated based on the enhanced speech using a non-linear transformation. Finally, the regenerated speech components are added into the Wiener filtered speech to obtain the improved speech. In this way, the over-attenuated speech components are recovered efficiently, and experimental results show clearly the good performance of our algorithm in achieving high quality speech enhancement.

2. WIENER FILTER

In the additive noise model, the noisy speech is the sum of the clean speech plus noise signal, which is also applicable in the frequency domain. If $Y(m,\omega)$, $S(m,\omega)$ and $N(m,\omega)$ represent the spectral magnitude of noisy speech, clean speech and noise signal respectively, and, θ_Y, θ_S and θ_N are their respective phases, the noisy speech can be expressed in the

time-frequency domain as

$$Y(m,\omega)e^{j\theta_Y} = S(m,\omega)e^{j\theta_S} + N(m,\omega)e^{j\theta_N} \quad (1)$$

where m is the frame index and ω is the frequency bin index.

In order to suppress the background noise while maintaining the speech content as much as possible, the traditional noise reduction algorithms try to obtain an estimated speech $\hat{S}(m,\omega)$ with minimal speech distortion. Generally, the estimated speech strongly relies on two important variables, the a-posteriori signal-to-noise ratio (SNR) $\gamma(m,\omega)$ and the a-priori SNR $\xi(m,\omega)$, which are defined respectively as follows:

$$\gamma(m,\omega) = \frac{Y(m,\omega)^2}{E\left[N(m,\omega)^2\right])} \quad (2)$$

$$\xi(m,\omega) = \frac{E\left[S(m,\omega)^2\right]}{E\left[N(m,\omega)^2\right])} \quad (3)$$

where $E[.]$ is the expectation function. In our paper, the background noise is assumed to be a stationary uncorrelated random process, and the expectation of the noise power is known. Therefore the a-posteriori SNR can be easily obtained. On the other hand, the a-priori SNR of each frequency bin can be calculated by the decision-directed approach [2] which is defined as

$$\hat{\xi}(m) = \alpha\frac{\hat{S}(m-1)^2}{E\left[N(m)^2\right]} + (1-\alpha)\max\left[\gamma(m)-1, 0\right] \quad (4)$$

where the frequency index ω is omitted for convenience and the parameter α is normally set to 0.98 for a good tradeoff between noise reduction and speech distortion. With estimated a-priori SNR $\hat{\xi}(m,\omega)$, the Wiener gain utilized in this paper can be expressed by Eq. (5).

$$g(m,\omega) = \frac{\hat{\xi}(m,\omega)}{1+\hat{\xi}(m,\omega)} \quad (5)$$

Finally the enhanced speech $\hat{S}(m,\omega)$ is obtained by

$$\hat{S}(m,\omega) = g(m,\omega)Y(m,\omega) \quad (6)$$

3. PROPOSED POST-PROCESSING TECHNIQUE

As discussed previously, the Wiener filtered speech $\hat{S}(m,\omega)$ suffers from distortions since some weak components of speech are considered as the background noise and are suppressed together with the noise by common noise reduction algorithms. In order to correct this problem, a post-processing technique is proposed.

3.1. Envelop Estimation

The objective is to regenerate the over-attenuated voiced and unvoiced speech components. Hence we choose the spectral subtraction filtered speech instead of the Wiener filtered speech $\hat{S}(m,\omega)$ to do the envelop estimation since the spectral subtraction causes lower distortion and it is without the one-frame delay problem which has been indicated in [8]. The estimated envelop $e(m,\omega)$ of a certain frame m_i can be expressed as

$$e(m_i,\omega) = \sqrt{\max\left[Y(m_i,\omega)^2 - N(m_i,\omega)^2, 0\right]} * H(\omega) \quad (7)$$

where $*$ is the convolution operator and $H(.)$ is a low-pass FIR filter with a cutoff frequency of around 150 Hz. By convoluting with a low-pass filter, a smoother and more continuous envelop is generated. Some weak speech components, such as the ones over-attenuated by the spectral subtraction algorithm can therefore be recovered by this step.

3.2. Excitation Generation

The non-linear transformation is a simple and efficient way to generate the excitation since it preserves the harmonic structure without any discontinuity in the spectrum [7]. The absolute value or full-wave rectification has been chosen instead of half-wave rectification as it generates a flatter spectrum. If the $\hat{s}(t)$ denotes the filtered speech $\hat{S}(\omega)$ in the time domain, the excitation signal $z(m,\omega)$ can be obtained as follows:

$$z(m,\omega) = \text{STFT}\{\text{abs}[\hat{s}(t)] * W(t)\} \quad (8)$$

where STFT(.) and abs(.) are the short-time Fourier transform and the absolute value function, respectively. The $W(.)$ is a whitening filter which is designed to flatten the spectrum of the full-wave rectified signal. This is done by performing a linear predictive coding (LPC) analysis and using the resulting coefficients to whiten the excitation. The reason for applying the whitening filter is to ensure that the regenerated speech will conform to the computed envelop better.

3.3. Speech Synthesis

It is well-known that in Wiener filtering, the energy of the filtered speech is lower than the energy of the clean speech. This can be partially reversed by synthesizing the final enhanced speech as a combination of filtered speech and a weighted portion of the artificially regenerated speech components. This process is given in Eq. (9).

$$\tilde{S}(m,\omega) = \hat{S}(m,\omega) + \beta \cdot z(m,\omega)e(m,\omega) \quad (9)$$

where the parameter β is empirically determined and equal to 0.1 in this paper. Unlike the method in [9] which low-passes the enhanced speech and just synthesizes the high frequency components, Eq. (9) is computed over all the frequency bands. It also preserves the contribution of the common noise reduction algorithms by including the enhanced speech $\hat{S}(m,\omega)$. The final regenerated speech harmonics are produced by the excitation signal $z(m,\omega)$ as well as the non-linear transformation used in [7]. Furthermore, some weak

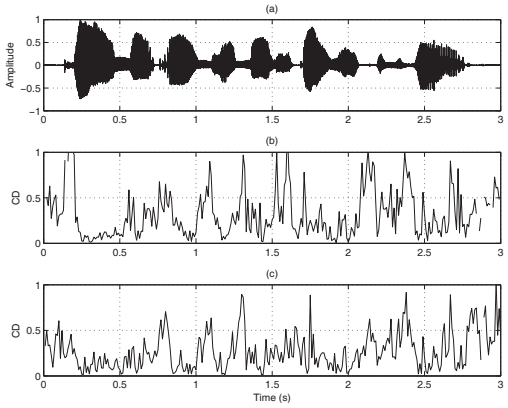


Fig. 1. (a)waveform of clean speech and CD between clean speech and (b) speech enhanced by HRNR (c) ours with white noise at SNR=5dB

spectral components, especially the unvoiced speech, could be recovered by the estimated envelop $e(m, \omega)$ which is extracted from the enhanced speech obtained using the spectral subtraction algorithm. Therefore the synthesized speech $\tilde{S}(m, \omega)$ is able to combine good noise reduction with lower speech distortion both for voiced and unvoiced speech components.

The synthesized speech $\tilde{S}(m, \omega)$ should undergo the inverse Fourier transform and the add & overlap process to produce the restored speech signal $\tilde{s}(t)$ in the time domain.

4. EXPERIMENTAL RESULTS

In our experiments, the proposed post-processing technique is evaluated by two objective measures, namely cepstral distance (CD) and perceptual evaluation of speech quality (PESQ). Ten utterances from the TIMIT database, spoken by 5 females and 5 males, are selected as the test data. Different noise types from the NOISEX database, including white noise, fan noise, car noise and f16 aircraft noise with various input SNRs ranging from -10 dB to 10 dB, are added. We compare our approach with the noisy speech (Noisy), the Wiener filtered speech (Wiener) and the recently proposed Harmonic Regeneration Noise Reduction (HRNR) approach [8]. The objective measures show that our method is better than the others, which is further confirmed by informal subjective listening tests. The proposed algorithm results in speech with more high frequency content and very little musical noise. This is because musical noise exists as isolated spectral peaks and they are masked to some extent by the regenerated speech components.

4.1. CD Measure

Cepstral distance can be used to measure the distance between two spectral envelops and it has been used in objective tests in [7, 8]. Fig. 1 presents the waveform of clean speech together with the CDs generated from HRNR and our approach. It clearly shows that although HRNR has a good performance for the voiced speech components, our post-processing technique regenerates the speech components for the entire speech including the unvoiced periods and hence achieves a better performance for high quality speech enhancement. The results also demonstrate the capability of our technique to recover both voiced and unvoiced speech components. The comparisons among Noisy, Wiener, HRNR and our technique based on the mean values of CD are shown in Table 1. The lower the value of CD, the better the performance of the corresponding algorithm. From the table, it can be seen that our algorithm achieves significant improvements especially for the low input SNR range.

Table 1. Mean Cepstral Distance Comparison

Noise type	SNR (dB)	Mean Cepstral Distance			
		Noisy	Wiener	HRNR	Ours
White	-10	2.93	2.32	3.15	**1.91**
	-5	2.73	1.99	2.52	**1.55**
	0	2.43	1.70	1.78	**1.28**
	5	2.06	1.36	1.29	**1.03**
	10	1.66	0.97	0.94	**0.78**
Fan	-10	2.01	1.57	1.60	**1.28**
	-5	1.58	1.06	1.16	**0.89**
	0	1.20	0.68	0.85	**0.58**
	5	0.88	0.42	0.62	**0.36**
	10	0.63	0.26	0.45	**0.23**
Car	-10	1.38	0.70	0.72	**0.60**
	-5	1.05	0.50	0.53	**0.45**
	0	0.80	0.36	0.41	**0.34**
	5	0.59	0.27	0.33	**0.26**
	10	0.44	0.19	0.27	**0.19**
F16 aircraft	-10	2.32	1.90	2.73	**1.49**
	-5	2.10	1.52	1.89	**1.24**
	0	1.77	1.14	1.30	**0.97**
	5	1.39	0.76	0.90	**0.68**
	10	1.02	0.46	0.66	**0.44**

4.2. PESQ Measure

The PESQ measure which aims to predict the results of subjective listening tests is described in ITU-T Recommendation P.862, and has been proven to be more reliable and correlated with Mean Opinion Score (MOS) than other traditional objective measures in most situations [10]. It yields an accurate evaluation both on speech distortion and noise distortion. The input SNRs of noisy speech and the corresponding PESQ results of Noisy, Wiener, HRNR and our technique are given in Fig 2. It can be found that our method is especially good under low input SNR conditions. These improvements become more obvious when the noise is mainly distributed in the low frequencies. Informal subjective results corroborated with the PESQ results.

5. CONCLUSION

Speech distortions caused by over-attenuation of speech components occurs in most traditional speech noise reduction algorithms since the weaker parts of speech are suppressed together with the noise. In this paper, a post-processing technique is proposed to alleviate this problem and it leads to high quality speech signal as a result. This is achieved by regeneration of both voiced and unvoiced speech components. The resulting speech is compared with the noisy speech, the Wiener filtered speech and HRNR filtered speech using two objective measures (CD and PESQ) and informal subjective listening tests. All the objective tests clearly show beyond doubt that the proposed algorithm is better and the improvements are especially noticeable at low input SNRs. Informal subjective listening tests indicate that the musical noise is suppressed and the restored speech sounds richer as it has more high frequency content.

6. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 113–120, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109–1121, 1984.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of IEEE*, vol. 67, pp. 1586–1604, 1979.

[4] I. Y. Soon and S. N. Koh, "Speech enhancement using 2-D Fourier transform," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 717–724, 2003.

[5] Y. Ephraim and D. Malah, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 251–266, 1995.

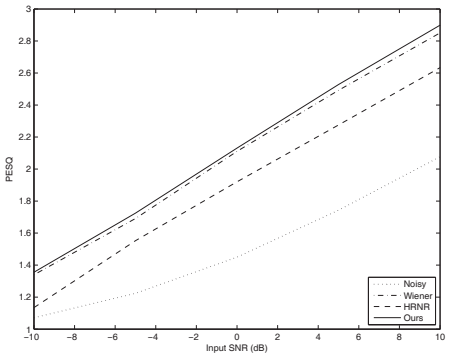
[6] Y. Hu and PC Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[7] C. Plapous, C. Marro, and P. Scalart, "Speech Enhancement Using Harmonic Regeneration," in *Proceedings ICASSP*, 2005.

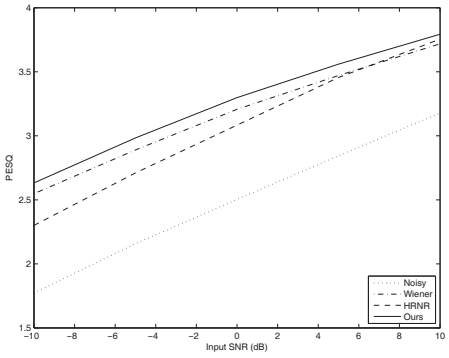
[8] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 2098–2108, 2006.

[9] X. Zhang, S.N. Koh, I.Y. Soon, and C. You, "Post-processing in masking-based β-order MMSE speech enhancement," *Applied Acoustics*, vol. 69, pp. 354–357, 2008.

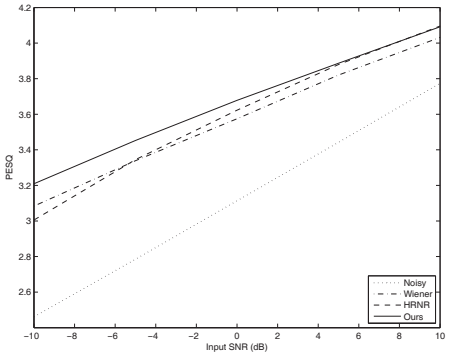
[10] Yi Hu and Philipos C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proceedings of INTERSPEECH-2006, Philadelphia, PA*, 2006.



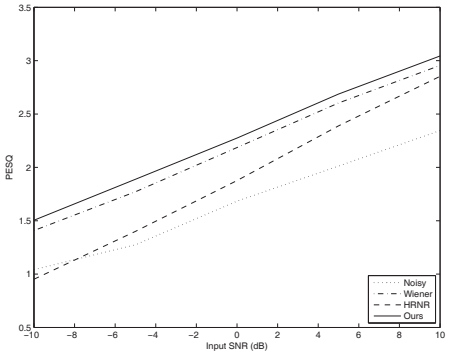
(a) Under white noise



(b) Under fan noise



(c) Under car noise



(d) Under f16 aircraft noise

Fig. 2. Comparison of PESQ results