Technical note

Post-processing in masking-based β-order MMSE speech enhancement

Xinxin Zhang *, Soo Ngee Koh, Ing Yann Soon, Changhuai You

Communication Research Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

Received 12 May 2006; received in revised form 26 November 2006; accepted 14 December 2006

Available online 27 February 2007

Abstract

A post-processing technique for improving the performance of the masking-based β-order MMSE method in speech enhancement is proposed in this letter. After the enhancement process, weak spectral components in the high-frequency band are usually over-attenuated. A non-linear high-frequency regeneration technique is used to re-synthesize the upper-band signal based on the lower-band speech.

Keywords: Speech enhancement; Noise reduction; β-Masking; MMSE

1. Introduction

In 1984, Ephraim and Malah derived a minimum mean-square-error (MMSE) short-time spectral amplitude (STSA) estimator [1], based on modeling speech and noise spectral components as statistically independent Gaussian random variables. The estimator was derived by minimizing the mean-square-error between the enhanced speech and the clean speech. Both noise and speech are modeled using the complex Gaussian model. You et al. subsequently proposed the adaptive β-order MMSE speech enhancement algorithm [2] and further incorporated masking properties in [3]. The β-order MMSE algorithm makes the filter gain \( G_β \) adaptive by adjusting \( β \) to a proper value in order not to over-attenuate the weak spectral components. \( G_β \) increases with the \( a \) priori SNR as the presence of speech is expected. For a given \( a \) priori SNR, \( G_β \) also increases as the \( a \) poste-riori SNR decreases. This is done so that weak spectral components with low SNR will be appropriately enhanced.

For a speech signal corrupted by a given noise, weak spectral components which are completely swamped by noise are very difficult to be recovered, thus reducing the chance of recovering the original harmonic structure. It is observed that for voiced frames, over-attenuation usually occurs in the high-frequency bands where the speech spectral components are relatively weak. It is a common problem to have speech components over-attenuated during the process of noise attenuation, even for the improved β-masking enhancement technique. Instead of recovering the original harmonic structure, slight tonal distortions are observed. In order to further improve the spectral quality of the enhanced speech, a post-processing method to achieve regeneration of harmonics is used.

The non-linear high-frequency regeneration technique (NHR) [4] is modified to regenerate the harmonics in the upper-band (2–4 kHz). It is slightly different from the original NHR scheme, which was implemented in the time domain. As the speech enhancement algorithm is performed in the frequency domain, the modified NHR is a hybrid frequency and time domain scheme as explained below.

2. Upper-band synthesis

Unlike the NHR scheme proposed in [4], frequency domain low-pass filtering (LPF) and high-pass filtering (HPF) are performed, as the β-masking gain, \( G_β \), is applied to the noisy spectrum in the frequency domain. It is therefore, more convenient to modify the spectral components after they have been enhanced by \( G_β \). The block diagram of the NHR method is shown in Fig. 1.
The β-masking enhanced speech, \( x(t) \), has an effective bandwidth of 4 kHz (i.e. 8 kHz sampling rate). The spectrum is denoted by \( X(k) \). For each frame of speech, voiced/unvoiced/silence decision is made. If an unvoiced or a silence frame is expected, no NHR operation will be applied. If a voiced frame is detected, the NHR technique will be applied. Firstly \( X(k) \) is low-pass-filtered to remove distorted tones in the high-frequency band. Next, the bandwidth of the filtered signal \( y_1(t) \) is doubled by using a full wave rectifier (absolute function in the time domain). We thus obtain \( y_2(t) \), which gives an initial estimate of the upper-band (2–4 kHz) signal. In order to avoid overlapping with the lower-band signal, \( y_2(t) \) is then high-pass filtered in the frequency domain to obtain the upper-band spectrum \( Y_3(k) \). The high-pass and low-pass filters correspond to linear-phase FIR filters with 21 coefficients each and 3 dB points at 2 kHz. In the frequency domain, the spectral amplitudes of the filters can be obtained from these coefficients. The lower-band spectrum \( Y_1(k) \) and the upper-band spectrum \( Y_3(k) \) are then added, resulting in the synthesized spectrum \( Y_4(k) \). Next, the output spectrum \( Y_{\text{out}}(k) \) is obtained by normalizing the energy level of \( Y_4(k) \) to that of the general speech spectral envelope. \( Y_{\text{out}}(k) \) thus inherits the harmonic structure of \( Y_4(k) \), while maintaining the spectral envelope (or the energy level) of \( X(k) \).

\[
Y_{\text{out}}(k) = Y_4(k) \frac{\text{envlp}(X(k))}{\text{envlp}(Y_4(k))}
\]

(1)

here the “envlp function” is as shown in Fig. 2.

3. Performance evaluation

Four female and four male speech signals (F1–F4 and M1–M4) were used in our simulation study. These speech signals were then corrupted to different segmental SNRs (−8.6 dB, −4.3 dB, 0 dB, and 4.3 dB) by adding Gaussian distributed white noises. The noisy speech signals were firstly enhanced by the β-masking method followed by post-processing using the proposed non-linear high-frequency regeneration technique (β-masking-NHR in short).

Subjective assessments show that the NHR post-processing method leads to significant improvement in the subjective quality of the final enhanced speech. In the β-masking enhanced speech, the distortions mainly reside in
the high-frequency band and sound like musical tones. In the proposed NHR-enhanced speech, the residual noises are suppressed to a very low level, without further distortion of speech quality. Fig. 3 shows the spectrograms of a voiced speech frame, for clean speech, unprocessed noisy speech, β-masking enhanced speech and β-masking-NHR enhanced speech, respectively. The harmonics of the original weak spectral components are masked in the noisy speech signals and are removed in the β-masking enhanced speech signals. What is more, residual noise can be observed in Fig. 3 (iii). As shown in Fig. 3 (iv), the NHR approach is able to regenerate the harmonics in the higher frequency band based on the information in the lower-band. By replacing the musical tone plus noise with speech harmonics, the perceptual quality of the enhanced speech is significantly improved.

Objective measurements also show the advantage of using NHR post-processing method. The use of SNR as a performance measure in the time domain is not appropriate in this case [4] as the phase values of the estimated high-frequency components will not match those of the original signal. Instead, the average log spectral distance (LSD) defined as follows is used as an objective performance measure:

\[
\text{LSD} = \frac{1}{M} \sum_{m=0}^{M-1} \left[ \frac{1}{\log \xi} \int_{-0.5 \omega_s}^{0.5 \omega_s} |\log_e|X(e^{j\omega})| - \log_e|Y(e^{j\omega})|\right]^2 d\omega \right]^{0.5}
\]

(2)

where \(M\) is the number of frames and \(\omega_s\) is the sampling frequency. The comparison between the LSDs of the two schemes is as shown in Table 1 which shows that the proposed β-masking-NHR method achieves lower average LSD distortion than β-masking method alone. The result is as expected because the proposed method aims to regenerate the masked spectral components.

In Table 1, the LSD result of 2.78 under row SG1 and column β-masking means that eight clean speech utterances (F1–F4, M1–M4) are added with noise with the resulting average segmental SNR of –8.6 dB, and they are then enhanced by the β-masking technique with the resulting LSD value of 2.78. The rest of the entries in the table are interpreted in the same way. Though it can be seen that the proposed NHR approach does lead to improvements in LSD values in general, the smaller average values still do not correlate well with the more significant subjective improvements perceived through listening tests. This reflects the fact that when the improvement in speech quality is due to very low energy spectral (in this case, high-frequency) components, the choice of objective quality measure is important in indicating the actual subjective quality of the processed speech. Therefore, another objective measure, namely the psycho-acoustically motivated distortion measure (PMD) [5], is used as an objective measurement for assessing the effectiveness of the proposed approach. The PMD distortion measure is defined as follows:

\[
\text{PMD}_{\text{seg}} = \frac{1}{M} \sum_{m=0}^{M-1} 10 \log_{10} \left( \frac{1}{w^2+1} \sum_{k=0}^{w^2+1} \left( X_k - \hat{Y}_k \right)^2 \right) \right) \right) \right) \right) \right)
\]

(3)

where \(w\) denotes the window size; \(X_k\) and \(\hat{Y}_k\) are the \(k\)th spectral components of the clean and processed speech signals, respectively. The comparison result is also tabulated...
in Table 1. It shows that the proposed β-masking-NHR enhanced speech signals have less spectral distortion than those enhanced by β-masking alone, and this is more consistent with the subjective improvement perceived through subjective listening tests.

4. Conclusion

A post-processing method based on NHR is used to regenerate the upper-band harmonic components of the voiced speech signals based on the lower-band harmonic information to improve the quality of β-masking enhanced speech. Subjective listening and spectrogram comparisons show that the proposed β-masking-NHR method can be used to regenerate the harmonics of the upper-band, and at the same time remove the residual musical tone in the β-masking enhanced speech signals. Objective comparisons based on log spectral distance and psycho-acoustically motivated distortion measurements show that the proposed method gives an overall improvement in speech quality as compared to the case which does not have NHR post-processing.

References