Musical Noise Reduction Based on Spectral Subtraction Combined with Wiener Filtering for Speech Communication

WANG Guang-yan*, ZHAO Xiao-qun, WANG Xia

*School of Information, Hebei University of Technology, Tianjin, China 300130 Email:lentilyy@126.com

Keywords: Speech Communication; Speech Enhancement; Spectral Subtraction; Wiener Filter; Musical Noise

Abstract

The goal of this paper is to propose a new technique for musical noise reduction used to alleviate some of the speech distortion introduced by the spectral subtraction (SS) process, particularly to eliminate the background musical noise of actual environment in speech communication or recognition system. The new speech enhancement approach combines spectral subtraction and the conventional Wiener filtering (CWF) in series connection to construct a two-stage hybrid system (named SS-CWF) in frequency domain to enhance the speech with additive musical noise. The noisy speech is recorded under the real background musical noise environment at a relatively lower signal-to-noise ratio. Simulation results of the proposed method, comparing with that of the conventional spectral subtraction, show better performance. The performance is evaluated by using the Log-Likelihood Ratio (LLR) measure, which is an objective evaluation measure based on linear predictive coding (LPC) techniques. Experiment results have shown that combination SS-CWF method is more robust and efficient. Meanwhile, the subjective evaluation results indicate that this method provides better speech quality with cleaner waveforms and spectrograms in time and frequency domain. Consequently, the proposed technique has complementary advantages of the spectral subtraction and Wiener filter.

1 Introduction

Voice quality and intelligibility are always important for communication systems. In order to obtain near-transparent speech communications, speech enhancement techniques have been employed to improve the quality and intelligibility of the noise corrupted speech. Therefore, speech enhancement has been a challenging topic of research for many years. The mainly purposes of speech enhancement is noise reduction. Approaches to retrieve enhanced speeches are plentiful, such as Wiener filtering [1], Kalman filtering [2], and spectral subtraction[3,4]etc. The algorithms, such as the wavelet transform [5], neural networks [6] and blind signal separation [7] etc, have also been proposed in the literature. In order to improve the robustness and efficiency of the speech enhancement method, it is feasible to combine two or three algorithms in different manners or in different domains. There are some approximately successful trials among the speech researchers [2,5,6,7].

The Wiener filtering and spectral subtraction type algorithms are widely used in speech enhancement field because of their low computational complexity and impressive performance. As all we known that the main weakness of the spectral subtraction method is the production of an annoying noise called musical noise, which is suffered from the over-subtracting of the spectral. However, the conventional Wiener filters have the characteristics of suppressing the noise frequencies with the other speech frequencies unchanging. The proposed method of this paper is directed toward combining the spectral subtraction with the Wiener filter to construct a hybrid speech enhancement method (called SS-CWF method) which in turn has the effect of reducing the musical noises whether produced by the subtraction-type algorithms or the real background music noises under different environment. Apart from being extremely annoying to the listeners, the musical noise also hampers the performance of the speech-coding algorithms to a great extent. Consequently, it is necessary to find a simple and effective noise-reduction approach to enhance the quality and intelligibility of speech signal. Moreover, it will benefit the speech communication system too.

2 Principle of the Method

The main objective of speech enhancement is to improve the perceptual aspects of speech such as overall quality or intelligibility. The method of spectral subtraction is closely related to the spectral weighting in a Wiener filter. Principles of the two signal processing approaches illustrate the feasibility of the proposed SS-CWF method.

2.1 The Wiener Filter

The Wiener filter rule is derived from the optimal filter theory. It is assume that the noisy signal is the clean speech $s(t)$ with the uncorrelated additive noise $n(t)$ and the noise obey normal distribution. The impulse response $h(t)$ of the filter is derived in the minimum mean squared error (MMSE) sense by minimizing the Euclidian distance $E\left\{\left|\hat{s}(t)-s(t)\right|^2\right\}$. Where $\hat{s}(t)$ is the estimate of the clean speech from the noisy speech $y(t)$, and $E\{\bullet\}$ denotes the expectation operator.

Assuming that both $s(t)$ and $n(t)$ are short-time stationary stochastic process, the essence is to solve the Wiener-Hopf equation

$$R_{sy}(\tau) = h(a)R_{yy}(\tau - a)da \qquad (1)$$

Perform the Fourier transform on the two sides of Eq. (1)

$$P_{sy}(\omega) = H(\omega)P_{yy}(\omega) \qquad (2)$$

The signals *s*(*t*) and *n*(*t*) are independent to each other, thus：

$$P_{yy}(\omega) = P_s(\omega) + P_n(\omega) \qquad (3)$$

Substitute Eq.(3) into Eq.(2), we get the frequency solution of Wiener-Hopf Equation

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_n(\omega)} \qquad (4)$$

Eq. (4) is the frequency response of the conventional Wiener filter. Where $P_s(\omega)$ and $P_n(\omega)$ is the power spectrum destiny (psd) of the speech and noise respectively. $P_n(\omega)$ can be estimated from the noisy speech during regions of silence by use of a noise estimation method, and $P_s(\omega)$ can be obtained from the input signal by use of the spectral subtraction method. In order that the finally estimated speech power spectra is given by applying the Wiener filtering on the original power spectra of noisy speech.

2.2 Idea of the Spectral Subtraction

The method of spectral subtraction is widely used for noisy speech enhancing. The basic idea of spectral subtraction is performed in the frequency domain by operating on the Fourier transformation of the observed samples. The estimation of the noise-reduced speech spectrum is obtained by subtracting an estimated mean spectral magnitude of the noise from the spectral magnitude of the noisy speech signal.

Let $y(k) = s(k) + n(k)$ be the sampled noisy consisting of the clean signal *s*(*k*) and the noise signal *n*(*k*). Firstly, the additive samples of the noisy speech signal should be windowed by a type of L-length window function with L/2 overlaps. The result signals denote as $y_W(k)$, $s_W(k)$ and $n_W(k)$ respectively. Then the additive model becomes

$$y_W(k) = s_W(k) + n_W(k) \qquad (5)$$

Taking the short-time Fourier transform of *y*(*k*), we get

$$Y(\omega) = S(\omega) + N(\omega) \qquad (6)$$

To obtain the short-time power spectrum of the noisy speech, we multiply $Y(\omega)$ in the above equation by its conjugate $Y^*(\omega)$. In doing so, Eq. (6) becomes

$$|Y(\omega)|^2 = |S(\omega)|^2 + |N(\omega)|^2 + 2\,\mathrm{Re}\left[S(\omega)N^*(\omega)\right] \quad (7)$$

where $N^*(\omega)$ is the conjugate of $N(\omega)$. The terms $|S(\omega)|^2$, $|N(\omega)|^2$ and $S(\omega)\cdot N^*(\omega)$ can not be obtained directly and are approximated as $E\{|S(\omega)|^2\}$, $E\{|N(\omega)|^2\}$ and $E\{S(\omega)\cdot N^*(\omega)\}$. Typically, $E\{|S(\omega)|^2\}$ is estimated during non-speech activity, and is denoted by $|\hat{S}(\omega)|^2$. If we assume that *n*(*k*) is zero mean and uncorrelated with the clean signal *s*(*k*), then the term $E\{S(\omega)\cdot N^*(\omega)\}$ reduce to zero [3].

Conventionally, from the above assumptions, the estimation of the clean speech power spectrum can be obtained by

$$|\hat{S}(\omega)|^2 = |Y(\omega)|^2 - E\{|N(\omega)|^2\} \qquad (8)$$

According to paper [4], in order to minimize the residual and musical noise, we introduce an over-subtraction factor α to modify Eq. (8)

$$|\hat{S}(\omega)|^2 = |Y(\omega)|^2 - \alpha\cdot E\{|N(\omega)|^2\},\ \alpha \ge 1 \qquad (9)$$

The resulting spectrum is down-limited at a minimum β level (the spectral floor):

$$|\hat{S}(\omega)|^2 = \begin{cases} |\hat{S}(\omega)|^2 & \text{if } |\hat{S}(\omega)|^2 > \beta\cdot E\{|N(\omega)|^2\} \\ \beta\cdot E\{|N(\omega)|^2\} & \text{otherwise} \end{cases} \qquad (10)$$

These modifications lead to minimizing the perception of the narrow spectral peaks by decreasing the spectral excursions and thus to lower the musical noise perception. The enhanced speech spectrum is obtained using the magnitude estimate $|\hat{S}(\omega)|$ of the enhanced speech and the phase $\theta_y(\omega)$ of the input noisy signal:

$$\hat{S}(\omega) = |\hat{S}(\omega)|\exp(j\theta_Y(\omega)) \qquad (11)$$

The phase of the input signal is used for reconstruction of the estimated signal spectrum based on the fact that the short-time spectral amplitude is more important for intelligibility and quality than the phase in human perception. This is presented by Wang and Lim in their work [8].

The over-subtraction method [4] assumes that the noise affects the speech spectrum uniformly and the over-subtraction factor subtracts an overestimate of the noise over the whole spectrum. However, that is not the case with real-world noise, especially for the real musical background noise. Thus, we present the following hybrid system to reduce the background musical noise in real environment.

3 Implementation and Results

3.1 The Hybrid Noise-reduction System

The main problem of spectral subtraction method is the production of musical noise. As well as the Wiener filter has the advantages of suppressing the frequencies where noise is present while remaining other frequencies unchanged. Thus, we present a musical noise reduction method to enhance the speech by series combining of the spectral subtraction and conventional Wiener filter in frequency domain. That is the so called hybrid SS-CWF speech enhancement method. The proposed hybrid system is illustrated by Fig. 1. It is a two-stage approach which is used to enhance the noisy speech in actual environment. The spectral subtraction is employed to suppress the stationary noise component approximately, and then the Wiener filter is employed to reduce the musical noise which generates from the first stage, along with the real background music. It tracks the clean speech and noise spectrum dynamically for each frame, which provides a first-stage estimation of clean speech power spectra and noise power spectra. These estimates are used to construct a Wiener filter for the original noisy speech power spectra, which removes noise while preserving speech spectra.

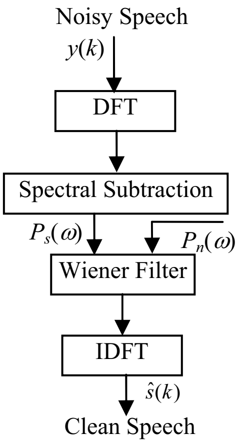
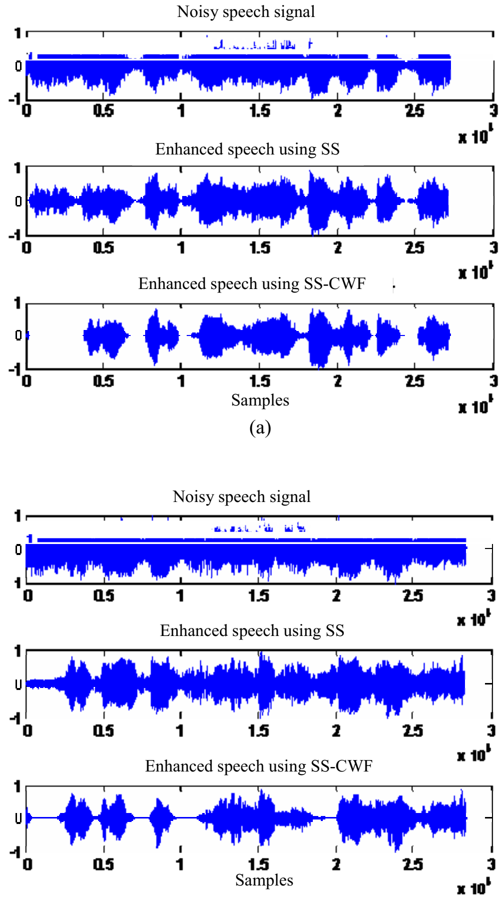


Fig. 1: Block diagram of the hybrid system



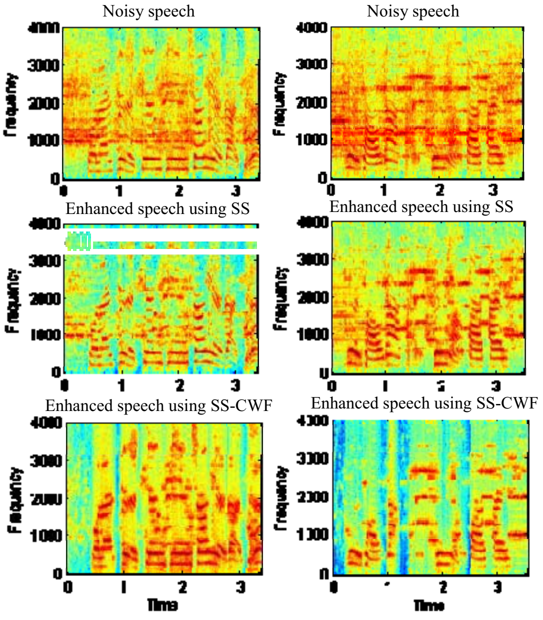
(a)

(b)

Fig. 2: Waveforms of the (a) Chinese utterance “wǒ lái zí Tiān jīn shāng yè dà xué” (b) Chinese utterance “miàn cháo dà hǎi, chūn nuǎn huā kāi”

3.2 Experiment Results

The original noisy speech signal comes from the database constructed by our research group. The speech database contains many terms of Chinese speech which recorded under the various controlled noise environments. Here, we take two pieces of sentence as the typical examples. The two test sentences are both Mandarin Chinese recorded by a female at a sampling frequency of 8-kHz, using 8-bit coding. The Chinese phonetic of the first one is “wǒ lái zí Tiān jīn shāng yè dà xué” whose background noise is a English song. The second is “miàn cháo dà hǎi, chūn nuǎn huā kāi” with a light music as the background noise. For the sake of convenience, the first sentence is denoted as speech A, and the second on is speech B. For comparative purposes, we extract the first-stage output from the spectral subtraction as the enhanced results of the conventional SS method. The Fig.2 and Fig.3 show the time evolutions (waveforms) and the spectrograms of speech A and B using the conventional SS method and the hybrid SS-CWF method respectively. Fig.(a) illustrates the results corresponding to speech A, and Fig.(b) corresponds to speech B. The result graph of each speech includes three sub-graphs arranged vertically from top to bottom. As follows in sequence， the sub-graphs illustrate the waveforms (in Fig.2) or spectrograms (in Fig.3) of the original noisy speech, the enhanced speech using the spectral subtraction method only, and the enhanced speech using the proposed SS-CWF method.



(a)

(b)

Fig. 3: Spectrograms of the (a) Chinese utterance “wǒ lái zí Tiān jīn shāng yè dà xué” (b) Chinese utterance “miàn cháo dà hǎi, chūn nuǎn huā kāi”

According to Fig.2 and Fig.3, it is found that the enhanced speech signal by use of the hybrid SS-CWF method is much cleaner and less distortion than that of the SS method. Simultaneously, by hearing the enhanced speech signals of different methods, we can feel that the musical noise and the residual background noise of SS system obviously. However, the output of the hybrid SS-CWF system is very clean and is nearly no musical noise. Of cause, hearing is the most important way of subjective quality measures.

4 Quality Evaluation

Many objective and subjective speech quality measures have been proposed in the past to evaluate the quality of speech for speech enhancement algorithms [9].

One of the three types of LPC-based objective measures is used for this experiment. That is the Log-Likelihood Ratio (LLR) measure [10], which is defined as:

$$d_{LLR}(A,B) = \ln\left[\frac{A \cdot V \cdot A^T}{B \cdot V \cdot B^T}\right] \qquad (12)$$

where $A = \left[1 \quad -a_1 \quad -a_2 \Lambda \ -a_N\right]^T$ is the LPC vector of the original speech signal frame, $B = \left[1 \quad -b_1 \quad -b_2 \Lambda \ -b_N\right]^T$ is the LPC vector of the enhanced speech frame, and V is the autocorrelation matrix of the original speech signal. The segmental LLR values were limited in the range of [0,2] to further reduce the number of outliers[10]. In this experiment, A represent the LPC vector of the original noisy signal. Consequently, the bigger of the d_{LLR} means the better quality of the enhanced speech.

Table 1: the LLR measure results

	SS method	The SS-CWF method
Speech A	0.2209	1.6877
Speech B	0.3301	1.2633

Table 1 shows the LLR measure results obtained by the two Chinese utterances using different speech enhancement methods. We can notice that the hybrid SS-CWF method exhibits bigger values than those obtained with the traditional SS method.

5 Conclusions

This paper addresses the problem of noise reduction of additive musical background noise in speech based on the series combination of the conventional spectral subtraction and Wiener filter in frequency domain. The proposed SS-CWF speech enhancement method is a two-stage hybrid system. After applying conventional Weiner filter as the second-stage processing, simulations showed a better quality with less distortion for enhanced speech and that the musical noise was effectively masked. In another words, the proposed SS-CWF speech enhancement method reduces both the residual musical tones that appear in the case of conventional power spectral subtraction and the background songs/music that appear in the actual environment. The hybrid SS-CWF system can produce a better and cleaner enhanced speech signal with less musical noise.

References

[1] L. Akter, Md. Kamrul Hasan. “CROSSCORRELATION COMPENSATED WIENER FILTER FOR SPEECH ENHANCEMENT”, *ICASSP*, Ⅰ, pp. 457-460, (2006).

[2] Leandro Aureliano da Silva, M. B. Joaquim. “Noise reduction in biomedical speech signal processing based on time and frequency Kalman filtering combined with spectral subtraction”, *Computers and Electrical Engineering*, **34**, pp. 154-164, (2008).

[3] Yang Lu, Philipos C. Loizou. “A geometric approach to spectral subtraction”, *Speech Communication*, **50**, pp. 453-466, (2008).

[4] Radu Mihnea Udrea, Nicolae Vizireanu, Silviu Ciochina, Simona Halunga. “Nonlinear spectral subtraction method for colored noise reduction using multi-band Bark scale”, *Signal Processing*, **88**, pp. 1299-1303, (2008).

[5] T. Giilzow, A. Engelsberg, U. Heute. “Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement”, *Signal Processing*, **64**, pp. 5-19, (1998).

[6] Yu Shao, Chip-Hong Chang. “A Novel Hybrid Neuro-Wavelet System for Robust Speech Recognition”, *ISCAS*, pp.1852-1855, (2006).

[7] Kenichi Furuya, Akitoshi Kataoka. “Robust Speech Dereverberation Using Multichannel Blind Deconvolution With Spectral Subtraction”, *IEEE Transactions on audio, speech, and language processing*, **15(5)**, pp.1579-1591, (2007).

[8] D.L. Wang, J.S. Lim. “The unimportance of phase in speech enhancement”, *IEEE Trans. Acoust. Speech Signal Process*, **30 (4)**, pp. 679–681, (1982).

[9] J.H.L. Hansen, B. Pellom. “An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms”, *ICSLP-98: Inter. Conf. on Spoken Language Processing,* **7**, pp. 2819-2822, (1998).

[10] Yi Hu, Philipos C. Loizou. “Evaluation of Objective Quality Measures for Speech Enhancement”, *IEEE Transactions on audio, speech, and language processing*, **16(1)**, pp. 229-238, (2008).