Adaptive Filtering for Attenuating Musical Noise Caused by Spectral Subtraction

[a)] *Takahiro Murakami and* [b)] *Yoshihisa Ishida*

[a)]Department of Electrical and Electronic Engineering,
Tokyo University of Agriculture and Technology, Japan
[b)]Department of Electronics and Communications, Meiji University, Japan
[a)]tmrkm@cc.tuat.ac.jp
[b)]ishida@isc.meiji.ac.jp

Abstract

A method of alleviating processing distortion caused by spectral subtraction is presented. It is well known that the spectral subtraction introduces annoying artifacts, which are referred to as undesirable *musical noise*, in the enhanced speech. The enhancement quality of the spectral subtraction quite depends on the performance of reducing the musical noise. Our approach exploits an adaptive filter in order to eliminate such distortion. In the method, the enhanced speech obtained by the spectral subtraction is used as a reference signal of the adaptive filter. The proposed method utilizes the characteristic difference between the musical noise and speech, i.e., the majority of the frequency components consisting the musical noise have the duration shorter than those of speech. Therefore, when the convergence speed of the adaptive filter is slower than the lifetime of the musical noise but faster than that of speech, only speech components can be tracked by the filter while the musical noise components are attenuated. Simulation results show that the proposed method can efficiently reduce the musical noise and the enhancement quality is improved in comparison with the conventional spectral subtraction.

Index Terms: spectral subtraction, musical noise, adaptive filter.

1. Introduction

In the last few decades, various applications based on speech signals processing, for example, automatic speech recognizers, hands-free mobile telephony, and hearing aids, have made rapid progress. Since most applications are developed to give an optimum performance for clean speech, a speech enhancement technique is necessary in order to provide better utility when we use such applications in real environments.

Spectral subtraction is one of the methods for such a purpose in speech signal processing [1, 2, 3, 4]. In the spectral subtraction, noise reduction is achieved by subtracting a noise spectrum from a noisy signal in the frequency domain. Since the spectral subtraction has relatively simple structure and low computational tasks, this algorithm is widely used. In very low signal-to-noise ratio (SNR) environments, however, the spectral subtraction introduces annoying artifacts, which are referred to as *musical noise*, in the enhanced speech. This self-introduced distortion causes the significant deteriorations of speech quality such as intelligibility. Therefore, attenuation of the musical noise is one of the key issues of giving the better performance [2, 3].

The method presented in this paper exploits an adaptive filter in order to alleviate the musical noise. In the method, the enhanced speech obtained by the spectral subtraction is used as the reference signal of the adaptive filter. Our method utilizes the characteristic difference between the musical noise and speech, i.e., the majority of the frequency components consisting the musical noise have the duration shorter than those of speech. Therefore, when the convergence speed of the adaptive filter is slower than the lifetime of the musical noise but faster than that of speech, only speech components can be tracked adaptively while the musical noise components are attenuated.

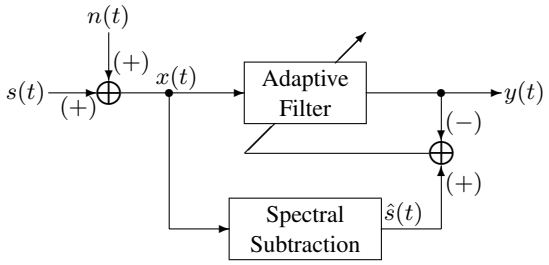

Figure 1: *A block diagram of the proposed speech enhancement system.*

2. Approach

Fig.1 illustrates a block diagram of the proposed speech enhancement system. In this figure, $s(t)$, $n(t)$, $x(t)$, $\hat{s}(t)$ and $y(t)$ are the clean speech, noise, noisy speech, enhanced speech obtained by the spectral subtraction, and the final output, respectively. The proposed method consists of two steps. First, the noise components in $x(t)$ is eliminated by using the conventional spectral subtraction and $\hat{s}(t)$ is obtained. Second, the adaptive filter is utilized for attenuating processing distortion caused by the spectral subtraction. As in Fig.1, $x(t)$ and $\hat{s}(t)$ are respectively used as the input and reference signals. Finally, $y(t)$ is given as the output of the adaptive filter.

2.1. Review of spectral subtraction and processing distortion

2.1.1. Spectral subtraction

The spectral subtraction technique is one of the methods of retrieving a signal of interest observed in additive noise. In the method, the power (or magnitude) spectrum of the original signal is esti-

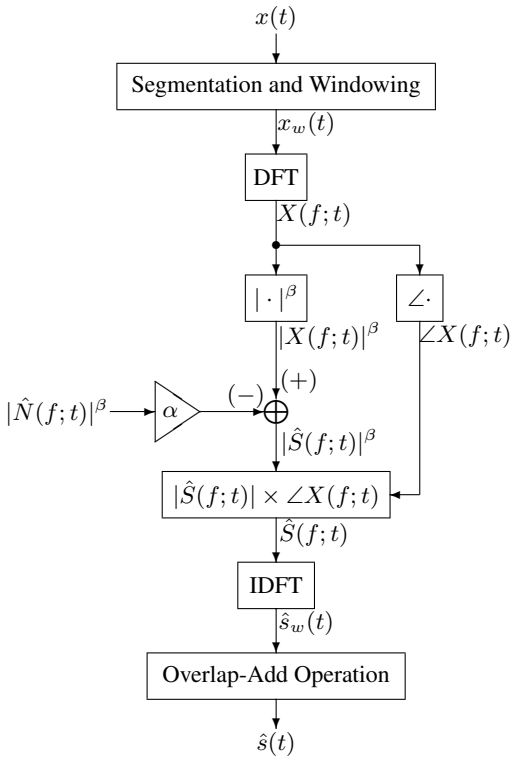

Figure 2: *A block diagram of spectral subtraction.*

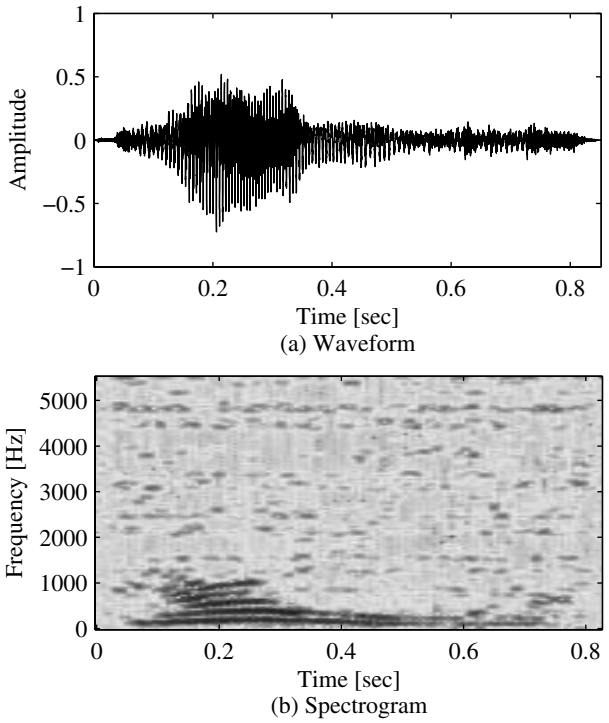

Figure 3: *A spectrogram of enhanced speech obtained by spectral subtraction (input SNR=0dB).*

mated from the observed noisy signal. Fig.2 summarizes a scheme of the spectral subtraction. As shown in the figure, the observed signal is transformed into the frequency domain via the discrete Fourier transform (DFT), and then the previously estimated noise spectrum is subtracted. Finally, the reconstructed time-domain signals is obtained by applying the inverse DFT (IDFT) to estimated power (or magnitude) spectrum combined with the phase of the observation. In terms of computational complexity, the spectral subtraction is relatively inexpensive. Therefore, this method is widely used in speech applications.

Let a noisy speech signal $x(t)$ be

$$x(t) = s(t) + n(t), \tag{1}$$

where $s(t)$ and $n(t)$ are the clean speech and noise signals, respectively. In the spectral subtraction, the observed noisy speech signal is buffered and divided into segments of L samples length, and each segment is windowed. The windowed signal can be expressed as

$$x_w(t) = s_w(t) + n_w(t), \tag{2}$$

In the frequency domain, Eq.(2) is rewritten by

$$X(f;t) = S(f;t) + N(f;t), \tag{3}$$

where $X(f;t)$, $S(f;t)$ and $N(f;t)$ are the f-th frequency components obtained by applying the DFT to $x_w(t)$, $s_w(t)$ and $n_w(t)$, respectively.

The spectral subtraction implements noise reduction by using only the power (or magnitude) spectra, i.e.,

$$|\hat{S}(f;t)|^\beta = |X(f;t)|^\beta - \alpha|\hat{N}(f;t)|^\beta, \tag{4}$$

where $|\hat{S}(f;t)|^\beta$ is the spectrum of enhanced speech, $|\hat{N}(f;t)|^\beta$ is the previously estimated noise spectrum, and α denotes the subtraction factor. For the power spectrum, the exponent β is set to $\beta = 2$ and for the magnitude spectrum, $\beta = 1$. Moreover, α is set to $\alpha > 1$ for the over-subtraction. The over-subtraction yields more improved results than the full noise subtraction, i.e., $\alpha = 1$. Estimated $|\hat{S}(f;t)|$ is combined with the phase spectrum of $X(f;t)$ as

$$\hat{S}(f;t) = |\hat{S}(f;t)| \times e^{j\angle X(f;t)}, \tag{5}$$

where $\angle X(f;t)$ is the phase spectrum of $X(f;t)$. Then, $\hat{S}(f;t)$ is transformed via the IDFT and the time-domain signal $\hat{s}_w(t)$ is obtained. Finally, $\hat{s}_w(t)$ is connected with the preceding and succeeding segments by the overlap-add operation to form the enhanced speech $\hat{s}(t)$.

In general, $|\hat{N}(f;t)|^\beta$ is estimated by using the signals of non-speech segments where speech is absent but only noise exists. An estimate of the noise spectrum is obtained as

$$|\hat{N}(f;t)|^\beta = \frac{1}{M}\sum_{m=1}^{M} |N(f;t_m)|^\beta, \tag{6}$$

where $|N(f;t_m)|^\beta$ is the spectra of the signals of the m-th non-speech segment.

2.1.2. Processing distortion caused by spectral subtraction

Since the spectral subtraction is based upon the underlying assumption that the statistics of noise process do not vary rapidly in time, the signals in the previous non-speech segments are available for the noise spectrum $|\hat{N}(f;t)|^\beta$ as in Eq.(6). $|\hat{S}(f;t)|^\beta$ obtained by Eq.(4), however, contains an estimation error between the estimated noise spectrum $|\hat{N}(f;t)|^\beta$ and the natural one $|N(f;t)|^\beta$. For instance, when the $|\hat{N}(f;t)|^\beta$ is greater than $|X(f;t)|^\beta$, the spectrum of the enhanced speech has a negative power (or magnitude) spectrum, in other words, the phase is constrainedly inverted. Especially in low signal-to-noise ratio (SNR) environments where SNRs are lower than 5dB, the spectral subtraction introduces undesirable metallic sounding noise, so-called *musical noise*. This distortion causes the significant deteriorations of the enhancement quality such as intelligibility.

The main cause of the musical noise is some frequency components which are randomly isolated, short-lived and narrow-bandwidth. Figs.3(a) and (b) respectively illustrate the waveform and the spectrogram of the enhanced speech obtained by applying the spectral subtraction. In this figure, the additive noise was assumed to be Gaussian and SNR=0dB. As shown, the isolated frequency components randomly appear at relatively high frequency region. The annoying musical noise is caused by these isolated frequency components.

The enhancement quality of the spectral subtraction quite depends on the performance of reducing the musical noise. In order to alleviate such distortion, the characteristic differences between musical noise and speech can be utilized. One of the most important characteristics of musical noise is that the majority of the frequency components consisting the musical noise have the duration shorter than about 20msec, whereas the duration of the speech components is considerably long [2]. However, precise distinction between musical noise and speech components are still difficult.

2.2. Attenuating musical noise using an adaptive filter

As shown in Fig.1, our approach exploits an adaptive filter in order to attenuate the musical noise. In the method, the noisy speech and the enhanced speech obtained by the spectral subtraction are used as the input and reference signals of the adaptive filter, respectively.

Let an input and output signals of an adaptive filter be $x(t)$ and $y(t)$, respectively. In the time domain, $y(t)$ is given by

$$\begin{aligned} y(t) &= \sum_{k=0}^{K} h(k;t)x(t-k) \\ &= h(t) * x(t), \end{aligned} \tag{7}$$

where $h(k;t)$ is the k-th coefficient of the adaptive filter at the time t and $*$ indicates the convolution operation. In the frequency domain,

$$Y(f;t) = H(f;t) * X(f;t), \tag{8}$$

where $Y(f;t)$, $H(f;t)$ and $X(f;t)$ are respectively the frequency components of $y(t)$, $h(k;t)$ and $x(t)$. The adaptive filter adjusts $h(k;t)$ (or $H(f;t)$) sample-by-sample so that a certain criterion is optimized. One of the criteria is to minimize an error between the reference, say $d(t)$, and output signals of the filter in the statistical sense. For instance, the normalized least-mean-square (N-LMS) algorithm is commonly used for the adaptive filter [5]. In the N-LMS algorithm, the mean of square error is minimized by updating its coefficients as follows:

$$\boldsymbol{h}(t+1) = \boldsymbol{h}(t) + \frac{\mu}{1+||\boldsymbol{x}(t)||^2}\boldsymbol{x}(t)\big(d(t) - \boldsymbol{x}^T\boldsymbol{h}(t)\big), \tag{9}$$

where $\boldsymbol{h}(t) = [h(0;t), h(1;t), \cdots, h(K;t)]^T$, $\boldsymbol{x}(t) = [x(t), x(t-1), \cdots, x(t-K)]^T$, $||\cdot||$ and T respectively indicate a norm and transpose of a vector, and μ $(0 < \mu \leq 1)$ denotes the step-size parameter which controls the convergence speed of the filter.

When the signals are stationary or those statistics vary slowly, the adaptive filter can track the statistics of the statistics. In the case of speech, if the noise is stationary, $X(f;t)$ can be regarded as stationary within short time-scale and the statistics of $X(f;t)$ vary slowly in a long range. Therefore, when the frequency components of $d(t)$, say $D(f;t)$, are stationary or vary slowly, the adaptive filter can tune $H(f;t)$ so that $Y(f;t)$ trails $D(f;t)$. On the other hand, the adaptive filter cannot track $D(f;t)$ whose statistics vary rapidly in comparison with the convergence speed of the filter. For example, when $D(f;t)$ has the relatively short duration, $Y(f;t)$ does not include such a frequency component. This is because $D(f;t)$ disappears before $H(f;t)$ converges.

As stated above, the lifetime of the frequency components of musical noise is considerably shorter than that of speech. Therefore, by using the speech containing the musical noise as the reference signal and setting the convergence speed of the adaptive filter so as to be faster than the duration of the speech components but slower than that of the musical components, only the musical noise can be filtered out.

3. Simulation

3.1. Parameter settings

In this simulation, the enhancement quality by the proposed method were compared with the cases of 1) the normal spectral subtraction (SS) and 2) the spectral subtraction with the post-processing (SS+PP) in which the musical noise is reduced by the conventional method, i.e., the frequency components having the duration shorter than 20msec are identified the musical noise and eliminated.

The speech used in this simulation were uttered by two male and two female speakers. The sampling frequency was 11.025kHz. The additive noise was assumed to be Gaussian and the additive noise level was varied from -5dB to 5dB.

In the spectral subtraction part, the length of analysis frame was set to $L = 256$, and the Hanning window was used. The subtraction factor was set to $\alpha = 2$. In this simulation, the magnitude spectral subtraction was employed, i.e., $\beta = 1$. Determination of the speech presence was achieved by manual inspection of the clean speech signals.

In the proposed method, the N-LMS algorithm was utilized for the adaptive filter. The length of the adaptive filter was empirically set to $K = 220$. To investigate the variation of the performance of the proposed method, the step-size parameter μ is varied to 0.1 (Proposed1), 0.5 (Proposed2), and 0.9 (Proposed3).

3.2. Performance measurements

For the objective measurement of the enhancement quality, in terms of both SNR [6] and averaged cepstral distance [7] were considered.

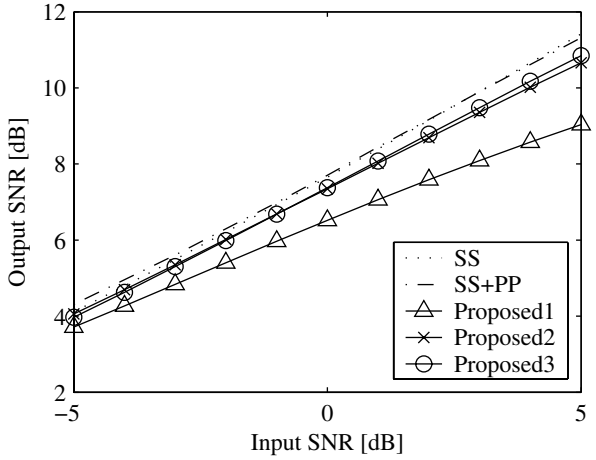

Figure 4: *Performance comparison in terms of SNR.*

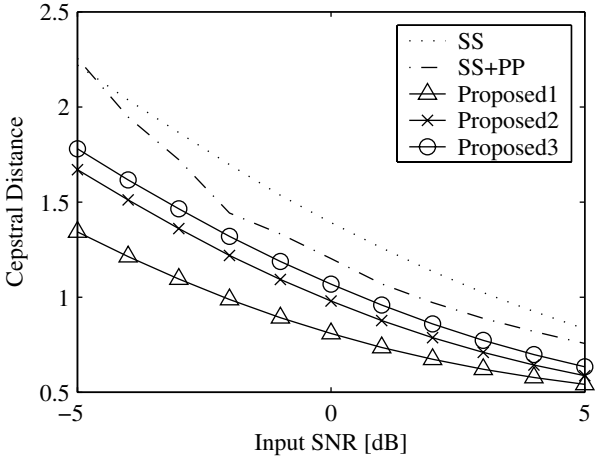

Figure 5: *Performance comparison in terms of averaged cepstral distance.*

SNR of the zero mean signal is calculated by

$$SNR(dB) = 10\log_{10}\left(\frac{\sum_t s^2(t)}{\sum_t \big(y(t)-s(t)\big)^2}\right), \qquad (10)$$

where P is the length of signals. On the other hand, the averaged cepstral distance is defined as

$$d_{cep} = \frac{1}{P}\sum_{p=1}^{P}\sum_{q=1}^{2Q}\big(c_s(q;t_p) - c_y(q;t_p)\big)^2, \qquad (11)$$

where P is the number of the frames where speech exists, Q is the order of the model, and $c_s(q;t_p)$ and $c_y(q;t_p)$ are respectively the cepstral coefficients corresponding to the clean and the enhanced speech at the p-th speech activity segment. d_{cep} is a non-negative index of intelligibility. As intelligibility is improved, d_{cep} decreases. In this simulation, Q was chosen as $Q = 8$, and the speech activity segment, i.e., t_p, are decided by manual inspection as the spectral subtraction part.

3.3. Simulation results

Figs.4 and 5 show the performance comparison in terms of SNR and averaged cepstral distance, respectively. In these figures, the performances were averaged over the four speech samples. As shown, Proposed3, i.e., the case where $\mu = 0.1$, greatly reduces the cepstral distance whereas SNR rather deteriorates in comparison with the cases of SS and SS+PP. On the other hand, as the step-size parameter increases, SNR is improved whereas the performance in terms of the averaged cepstral distance somewhat decays as compared with the Proposed3 case. Nevertheless, Proposed1 and Proposed2 give more improved performance in terms of the averaged cepstral distance than that of SS and SS+PP. Especially in the case of Proposed2, i.e., $\mu = 0.5$, the averaged cepstral distance is efficiently eliminated while SNR is maintained. Therefore, it can be said that by adjusting the step-size parameter adequately, the enhancement quality is efficiently improved by the proposed method.

4. Conclusion

A scheme for alleviating the effects of the musical noise caused by the spectral subtraction has been proposed. In the method, the enhanced speech by the conventional spectral subtraction is used as the reference signal of the adaptive filter. Simulation results have been shown that by exploiting the convergence property of the adaptive filter, the musical noise is effectively attenuated. Future works include to investigate the theoretically reasonable step-size parameter.

5. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol.ASSP-27, No.2, pp.113-120, April 1979.

[2] S. V. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, Second Edition, John Wiley & Sons, 2000.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. on Speech, Audio Processing, Vol.9, No.5, pp.484-487, July 2001.

[4] H. Nakashima, Y. Shicaki, T. Usagawa, and M. Ebata, "Spectral subtraction based on statistical criteria of the spectral distribution," IEICE Trans. Fundamentals, Vol.E85-A, No.10, pp.2283-2292, October 2002.

[5] S. Haykin, Adaptive Filter Theory, Fourth Edition, Prentice Hall, New Jersey, 2002.

[6] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, Discrete-Time Processing of Speech Signal, Back in Print, John Wiley & Sons, New York, 2000.

[7] R. L. B. Jeannés, A. Akbari Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," IEEE Trans. Speech, Audio Processing, Vol.5, No.5, pp.484-487, September 1997.