SPECTRAL SUBTRACTION BASED ON SPEECH/NOISE-DOMINANT CLASSIFICATION

Yukihiro NOMURA†, *Jianming LU*†, *Hiroo SEKIYA*†, *and Takashi YAHAGI*†

† Graduate School of Science and Technology, Chiba University
1-33, Yayoi-cho, Inage-ku, Chiba, 263-8522 Japan
ynomura@graduate.chiba-u.jp

ABSTRACT

This paper presents a spectral subtraction using the classifications between the speech dominant and the noise one. In our system, a new classification scheme between the speech dominant and the noise one is proposed. The proposed classifications use the standard deviation of the spectrum of observation signal in each critical band. We have introduced two oversubtraction factors for speech dominant and noise one, respectively. And spectral subtraction is carried out after the classification. The proposed method is tested on several noise types from the Noisex-92 database. On the basis of segmental SNR, inspection of spectrograms and listening tests, the proposed system is shown to be effective to reduce background noise. Moreover, our system generates less musical noise and distortion than the conventional systems.

1. INTRODUCTION

It is a fundamental and important problem to reduce the noise in the field of audio signal processing, when the speech signal is corrupted by additive background noise. Spectral subtraction [1]-[5] is a traditional approach for reducing background noise in single channel systems. And it is popular since it can suppress noise effectively, even in some real-life scenarios.

Recently, the spectral subtraction without assumptions about noise statistics was proposed[5]. In this method, the decisions on the speech dominant and the noise one are carried out in each critical band[1] that consists of several frequency bands. According to the decision, estimation of noise spectrum is carried out based on both the result of the decision and the masking[2] property of the human auditory system. And the noise is reduced using spectral subtraction. However, the enhanced speech includes musical noise and distortion, especially at low SNR. At low SNR, the incorrect classification of dominants occurs since the amount of noise is increased. Therefore, the error arises in the estimation of the noise spectrum, and the enhanced speech includes musical noise and speech distortion. Hence, it is required that the classifications between the speech dominant and the noise one are realized with reduced musical noise and speech distortion.

In this paper, we propose a new classification scheme between the speech dominant and the noise one. The proposed classifications use the standard deviation of the spectrum of observation signal in each critical band. If there are speech and noise components in a critical band, standard deviation is high. On the other hand, if only noise component is present in a critical band, standard deviation is low. Therefore, speech/noise-dominant can be classified by setting suitable threshold. We have introduced two oversubtraction factors for speech dominant and noise one, respectively. And spectral subtraction is carried out after the classification. The proposed method also requires no assumption about noise statistics. Moreover, the proposed method classifies the speech dominant and the noise one more correctly than the conventional method since the standard deviation is less influenced of by noise than the classification of conventional method. Therefore, the error is reduced in the estimation of the noise spectrum, and the enhanced speech using proposed method achieves less musical noise and speech distortion. We show the performance evaluation with spectrograms, segmental SNR and listening tests, and illustrate the effectiveness of the proposed system for reducing background noise.

2. CONVENTIONAL METHOD

2.1. Spectral subtraction [3]

We consider a speech signal $s(k)$ corrupted by an additive background noise $n(k)$. The observation signal $y(k)$ can be expressed by

$$y(k) = s(k) + n(k) \quad (1)$$

$$Y(\omega, r) = S(\omega, r) + N(\omega, r) \quad (2)$$

where $Y(\omega, r)$, $S(\omega, r)$ and $N(\omega, r)$ denote the short-time Fourier transforms of $y(k)$, $s(k)$ and $n(k)$ for frame r, respectively. Also, $s(k)$ is assumed to be uncorrelated with $n(k)$. If the noise spectrum $|N(\omega, r)|$ is estimated as $|\hat{N}(\omega, r)|$,

[1] For a given frequency, the critical band is the smallest band of frequencies around it which activate the same part of the basilar membrane in the ear.

[2] Masking is a process where one sound is rendered inaudible due to the process of another sound.

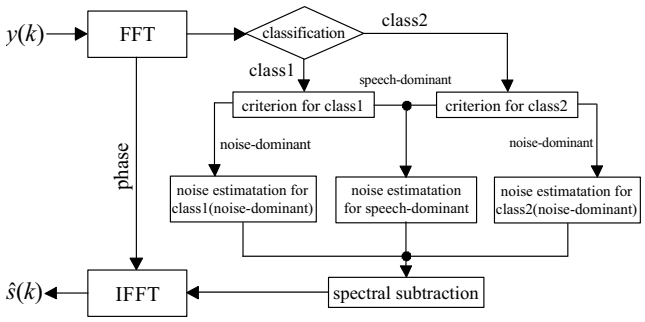

Fig. 1. Overall flow of the method [5]

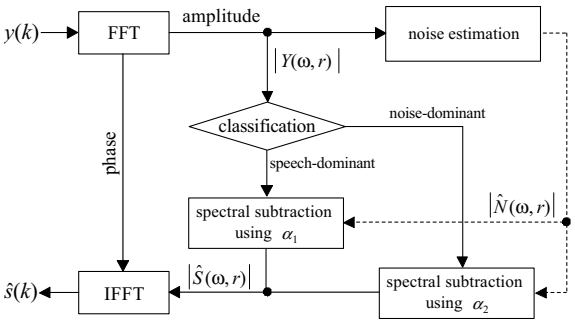

Fig. 2. Overall flow of the proposed method

the estimation of the short-time speech spectrum $|\hat{S}(\omega, r)|$ is represented by

$$|\hat{S}(\omega, r)| = H(\omega, r)|Y(\omega, r)| \tag{3}$$

$$H(\omega, r) = \begin{cases} \sqrt{1 - \alpha\, SNR_{post}(\omega, r)^2} & \text{if } J \geq 0 \\ \sqrt{\beta\, SNR_{post}(\omega, r)^2} & \text{otherwise} \end{cases} \tag{4}$$

$$J = \frac{1}{\alpha + \beta} - SNR_{post}(\omega, r)^2 \tag{5}$$

$$SNR_{post}(\omega, r) = |\hat{N}(\omega, r)|/|Y(\omega, r)| \tag{6}$$

where $H(\omega, r)$ is gain function, $\alpha(\geq 1)$is the oversubtraction factor and $\beta(\geq 0)$ is flooring level factor. When $J \geq 0$, spectral subtraction is carried out. On the other hand, spectral flooring is carried out when $J < 0$.

Once the subtraction is calculated in the spectral domain with (3) and (4) the enhanced speech signal $\hat{s}(k)$ is obtained as

$$\hat{s}(k) = IFFT\left[|\hat{S}(\omega, r)| \cdot e^{j \arg(Y(\omega, r))}\right] \tag{7}$$

where the phase of the observation signal is used for the enhanced speech signal.

2.2. Speech Enhancement Based on Speech/Noise-Dominant Decision

Figure 1 shows the overall flow of the method [5]. This method enhances speech in three steps.

The observation signal $y(k)$ is windowed by Hamming window, and the windowed signal is transformed into the frequency domain using FFT. The spectra of the observation signal pertaining to each critical band are summed, and the sums in each critical band for the previous L frames are arranged in ascending order. Next, the ordered sequence is classified into two classes using an approximation function. In the case of class 1 of Fig.1, a considerable quantity of speech component is present in the last L frames. On the other hand, in the case of class 2 of Fig.1, only noise and a small amount of speech component are present in the last L frames. Each class has a different criterion for evaluating the state of a particular band in a frame if it is speech or noise dominant. At last, estimation of noise spectrum based on [4] is carried out using the result of the decision and the masking property of the human auditory system. And spectral subtraction is carried out.

However, the enhanced speech includes musical noise and distortion, especially at low SNR. At low SNR, the incorrect classification of dominants occurs since the amount of noise is increased. Therefore, the error arises in the estimation of the noise spectrum, and the enhanced speech includes musical noise and speech distortion. Hence, it is required to realize the accurate classifications between the speech dominant and the noise one.

3. PROPOSED METHOD

3.1. Overview

Figure 2 shows the overall flow of the proposed method. The proposed method is composed of the following main steps:

1) The observation signal $y(k)$ is windowed by Hamming window, then the windowed signal is transformed to the frequency domain by applying FFT.
2) The classifications of speech/noise-dominant using the spectrum of observation signal are carried out. This step is described in detail in the next subsection.
3) Estimation of noise spectrum is carried out in the same manner as [4] .
4) After the classification, spectral subtraction based on (3) and (4) is carried out. We set two oversubtraction factors for speech dominant and noise one, namely α_1 and α_2. α_1 is set as the value to reduce the noise with little distortion to speech. α_2 is set $\alpha_1 < \alpha_2$ to fully reduce the noise.
5) By combining the $|\hat{S}(\omega, r)|$ and the phase of the observation signal, the enhanced speech signal $\hat{s}(k)$ is obtained by applying the inverse FFT.

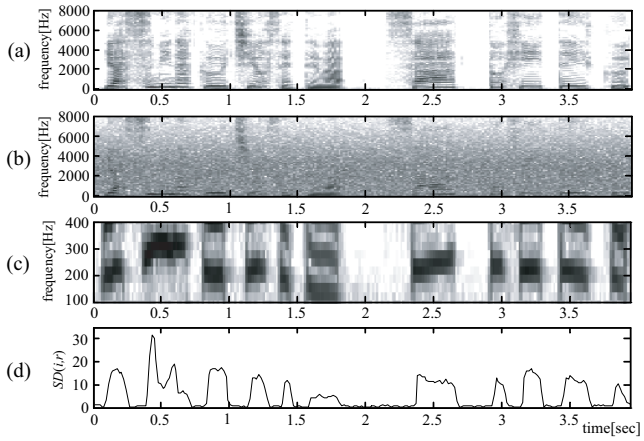

Fig. 3. Example of speech/noise-dominant classification using the spectrum of observation signal (a) Spectrogram of clean speech, (b) Spectrogram of observation signal (additive white noise with SNR=5dB), (c) Spectrogram of observation signal (range : 100-400Hz), (d) The output of $SD(i,r)$ ($i = 3$)

3.2. Speech/noise-dominant classification

We propose a new classification scheme between the speech dominant and the noise one. The proposed classifications use the standard deviation of the spectrum of observation signal in each critical band.

In Fig. 3 the spectrograms of (a) clean speech, (b) observation signal degraded by additive white noise with SNR=5dB and (c) noisy speech range from 100 to 400 Hz are shown. The spectrum of an observation signal is considered in a small region called critical band. From Fig. 3(b) and Fig. 3(c), the speech component is characterized by dark parallel stripes while the noise component is characterized by gray patches. Therefore, the standard deviation of the spectrum of observation is calculated. If speech and noise components are present in a critical band, the value of standard deviation is high. On the other hand, if only noise component is present in a critical band, the value of standard deviation is reasonably low. Therefore, speech/noise-dominant will be classified by setting a suitable threshold.

Figure 3(d) shows the output of $SD(i,r)$ in the 3-rd critical band (frequency:187-312Hz). The standard deviation of the spectrum of observation signal is calculated by

$$SD(i,r) = \sqrt{\sum_{\omega \in CB_i} \left\{ |Y(\omega,r)| - \overline{|Y(i,r)|} \right\}^2} \quad (8)$$

$$\overline{|Y(i,r)|} = \frac{1}{B_i} \sum_{\omega \in CB_i} |Y(\omega,r)| \quad (9)$$

where i is the number of critical band, CB_i is a set of frequency bands belonging to the i-th critical band, and B_i is the number of frequency bands in CB_i. From Fig. 3, the region with the speech component in Fig. 3(a) and Fig. 3(b) correspond to the high standard deviation in Fig. 3(d). Therefore, speech/noise-dominant can be classified by setting a suitable threshold Th as follows.

$$\begin{cases} \text{speech-dominant} & \text{if } SD(i,r) > Th \\ \text{noise-dominant} & \text{otherwise} \end{cases} \quad (10)$$

The proposed classification is more accurate than the conventional method since the standard deviation is less influenced of noise than the conventional method. Therefore, the error is reduced in the estimation of the noise spectrum, and the enhanced speech using the proposed method contains less musical noise and speech distortion. We show the performance evaluation with spectrograms, segmental SNR and listening tests, and illustrate the effectiveness of the proposed system for reducing background noise.

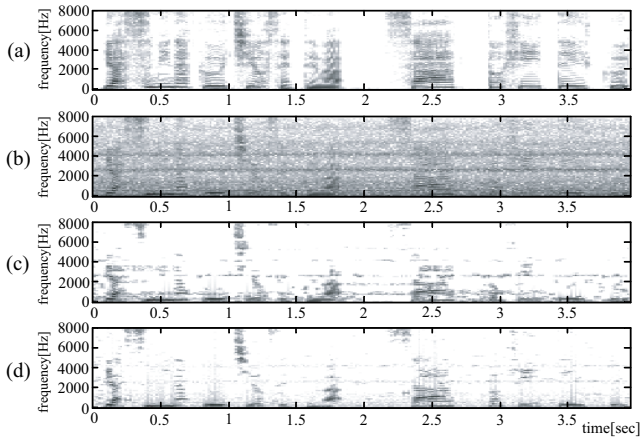

Fig. 4. Spectrograms of (a) Clean speech, (b) Observation signal(additive F16 noise with SNR=5dB), (c) Enhanced speech by the method [5], (d) Enhanced speech by the proposed method

4. PERFORMANCE EVALUATION

This section presents the performance evaluation of the proposed method. The following parameters have been chosen: 1) Hamming window of length $N = 512$ (32ms) with 50 % overlap; 2) $\alpha_1 = 2.5$, $\alpha_2 = 6$, $\beta = 0.001$, $Th = 5$, $q = 0.7$.

Four different background noises for this evaluation are taken from the Noisex-92 database. The four noises are white noise, pink noise, babble noise and F16 noise. As a speech signal, the Japanese female voice, */yasu mi naku uchi yosete wa saa tto hi-ite iku/*, is adopted. Noise is added to the clean speech signal with various SNR. The proposed method is compared with the methods [4], [5].

In Fig. 4, the spectrograms of (a) noisy speech, (b) degraded by F16 noise with SNR=5dB, (c) enhanced speech by the method [5], and (d) enhanced speech by the proposed method, are shown. From Fig. 4(c), we see that the noise

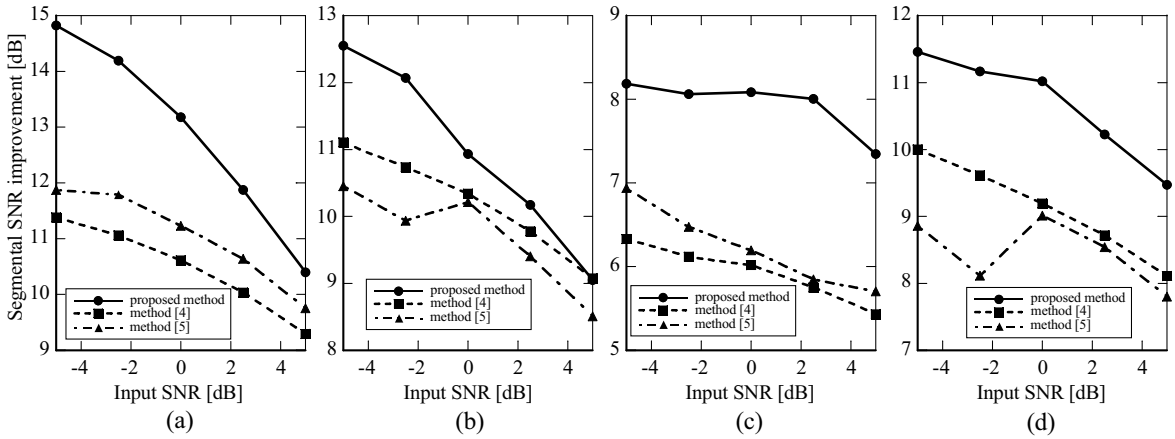

Fig. 5. Segmental SNR improvement for (a) White noise, (b) Pink noise, (c) Babble noise, (d) F16 noise

is reduced, but much isolated short stripes called musical noise are generated. On the other hand, the enhanced speech by the proposed method contains less musical noise than the method [5] as shown in Fig. 4(d).

In Fig. 5, the improvement of segmental SNR for (a) white noise, (b) pink noise, (c) babble noise and (d) F16 noise are shown. The segmental SNR is defined by

$$seg.SNR = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{k=Nm}^{Nm+N-1}s^2(k)}{\sum_{k=Nm}^{Nm+N-1}\{\hat{s}(k)-s(k)\}^2} \quad (11)$$

where N is the segment length and M is the number of segments in the speech signal. From Fig. 5, we find that the proposed method is superior to the conventional methods [4],[5] since the proposed method reduces the musical noise.

In addition to segmental SNR which is an objective measure, Mean Opinion Score (MOS) test is also performed to evaluate the performance of the proposed method. About 10 listeners are chosen randomly for each evaluation on noisy speech data. After listening to the original clean speech and noisy speech, each listener listens 3 times to each processed speech data, which are prepared in a random order. And the scores obtained from the listeners for each speech data are averaged. Table 1 shows the results of MOS tests for SNR=5dB. From Table 1, we find that the proposed method is better than the methods [4],[5] since the proposed method reduces the background noise with less musical noise and speech distortion. From the results of spectrograms, segmental SNR and listening tests, we find that the proposed system is effective to reduce background noise. Moreover, the proposed system generates less musical noise and distortion than the conventional systems.

5. CONCLUSION

In this paper, we have proposed a new classification scheme between the speech dominant and the noise one. The proposed classifications use the standard deviation of the spectrum of observation signal in each critical band. We have introduced two oversubtraction factors for speech dominant and noise one, respectively. And spectral subtraction is carried out after the classification. From the results of spectrograms, segmental SNR and listening tests, we find that the proposed system is effective to reduce background noise. Moreover, the proposed system generates less musical noise and distortion than the conventional systems.

Table 1. Comparison of MOS tests for SNR=5dB

	white	pink	babble	F16
proposed method	3.03	3.10	2.86	3.43
method [4]	2.90	2.83	2.70	2.83
method [5]	2.96	2.86	2.80	3.06

6. REFERENCES

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

[2] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and projection, for robust recognition in cars", Speech Commun., vol. 11, pp. 215-28, June 1992.

[3] H. Nakashima, Y. Chisaki and T. Usagawa, "Spectral subtraction based on statistical criteria of the spectral distribution", IEICE Trans. Fundamentals, vol.E85-A, no.10, pp.2283-2292, Oct. 2002.

[4] V. Stahl, A. Fischer and R. Bippusuchi, "Quantile based noise estimation for spectral subtraction and Wiener filtering ", in Proc. IEEE ICASSP, pp.1875 -1878, Istanbul, Turkey, Jun. 2000.

[5] S. Yoon and C. D. Yoo, "Speech enhancement based on speech/noise-dominant decision", IEICE Trans. Inf. & Syst., vol.E85-D, no.4, pp.744-750, Apr. 2002.