LETTER

Noise Estimation Using High Frequency Regions for Spectral Subtraction

Junpei YAMAUCHI[†a)], ***Student Member and*** **Tetsuya SHIMAMURA**[†], ***Regular Member***

SUMMARY This paper presents an improved spectral subtraction method for speech enhancement. A new noise estimation method is derived in which the noise is assumed to be white. By using the property that a white noise spectrum is flat, high frequency components of a noisy speech spectrum are averaged and the standard deviation of the noise is estimated. This operation is performed in the analysis segment, thus the spectral subtraction method combined with the new noise estimation method does not need non-speech segments and as a result can adapt to non-stationary noise conditions. The effectiveness of the proposed spectral subtraction method is confirmed by experiments.

key words: *spectral subtraction, noise estimation, high-frequency regions, non-stationary noise condition*

1. Introduction

Background noise acoustically added to speech can degrade the performance of speech processing systems. For this reason, various speech enhancement methods such as spectral subtraction, adaptive filter, correlation function, and the use of a speech model have been proposed up to now [1]. In a number of the speech enhancement methods, it is desirable to know a prior the standard deviation (or variance) of the noise. This is because in certain applications the noise is assumed to be white, and if its standard deviation or equivalently spectrum is known, then it is possible to eliminate the noise components from the noisy speech signal. The most popular speech enhancement method utilizing this may be the spectral subtraction based one [2].

Among some variations of the spectral subtraction based method, Boll's method [3] is well known and has been widely used. The use of Boll's method, however, is essentially restricted, because the noise must be estimated from some non-speech segments preceding the speech segment in this method. If the noise is stationary, then the estimated noise is accurate and Boll's method becomes fruitful. However, in practice, the noise is non-stationary in many cases, in which Boll's method may provide a degraded performance. To combat such a problem, Paliwal [4] proposed a noise estimation method based on the high-order Yule-Walker equations and demonstrated the performance of the spectral subtraction method combined with the estimation method. In Paliwal's method, the variance of the white noise can be estimated from only the analysis segment of the noisy speech signal, without using non-speech segments. Thus Paliwal's method can adapt to non-stationary noise conditions. However, this method needs a large amount of computation and suffers from numerical unstability.

In this paper, we derive a new noise estimation method and propose a spectral subtraction method combined with the new estimation method. By estimating the white noise components from high-frequency regions of the noisy speech spectrum, the proposed method can also adapt to non-stationary noise conditions. The amount of computation required for the proposed method is, however, extremely small compared with that required for Paliwal's method. And, the proposed method is numerically robust.

2. Spectral Subtraction

We assume that speech and noise are additive for noisy speech. Thus the noisy speech signal is, in the time domain, given by

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

where $s(k)$ and $n(k)$ are the speech signal and additive noise, respectively. In the frequency domain, the noisy speech signal of Eq. (1) is expressed as

$$X(f) = S(f) + N(f) \tag{2}$$

where $X(f)$, $S(f)$ and $N(f)$ are the Fourier transform of $x(k)$, $s(k)$ and $n(k)$, respectively.

From Eq. (2), the equation describing the spectral subtraction may be expressed as

$$|\hat{S}(f)| = \begin{cases} |X(f)| - \beta|N(f)| \\ \qquad\qquad |X(f)| - \beta|N(f)| \geq 0 \\ 0 \qquad\quad |X(f)| - \beta|N(f)| < 0 \end{cases} \tag{3}$$

where $\hat{S}(f)$ is an estimate of the original speech spectrum $|S(f)|$ and β is an over-subtracting factor. In this work, we use $\beta = 1$ for simplicity. The noise spectrum $|N(f)|$ is, generally, estimated from some non-speech segments by averaging the spectrum at each segment.

Manuscript received July 12, 2001.
Manuscript revised October 24, 2001.
Final manuscript received November 26, 2001.

[†]The authors are with the Department of Information and Computer Sciences, Saitama University, Saitama-shi, 338-8570 Japan.
a) E-mail: junpei@sie.ics.saitama-u.ac.jp

For restoration of the speech signal $s(k)$, the speech spectrum estimate $|\hat{S}(f)|$ is combined with the phase of the noisy speech signal, and transformed into the time domain via the inverse discrete-time Fourier transform as

$$\hat{s}(k) = IDFT[|\hat{S}(f)|e^{j\theta(f)}] \tag{4}$$

where $\theta(f)$ is the phase characteristic of $X(f)$.

3. Proposed Method

3.1 Noise Estimation

The new noise estimation method uses high-frequency regions of the noisy speech spectrum, which are more than 10 kHz. This is because it is assumed that the noise is white in this work. A white noise spectrum is flat, and as shown in Fig. 1 as an example a speech spectrum is almost absent in high-frequency regions. These directly lead to the fact that only the white noise spectrum appears in high-frequency regions, if we use a high sampling frequency (more than 20 kHz) to digitalize the noisy speech. Based on this principle, in the proposed method, high frequency components of the spectrum of noisy speech digitalized with a high sampling frequency are averaged, and the standard deviation (square root of the variance) of the white noise is evaluated. Even if the noise is non-stationary, this noise estimation approach may not degrade the performance of the combined spectral subtraction method, because the noise is estimated in the analysis segment of the noisy speech signal.

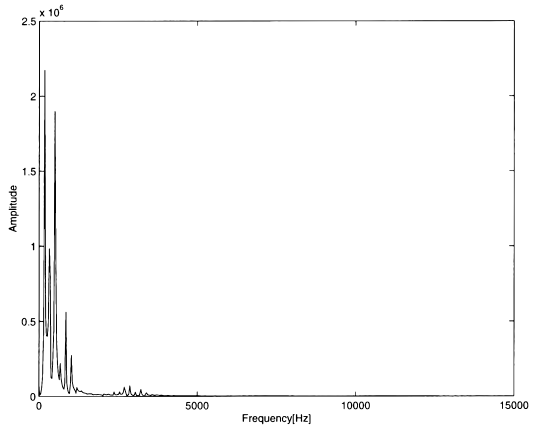

Fig. 1 Spectrum of a male vowel /a/.

3.2 Implementation of Spectral Subtraction

The procedure of the spectral subtraction method proposed in this paper is as follows. A block diagram of the method is drawn in Fig. 2.

1. A noisy speech signal segmented by Hamming window is transformed into the frequency domain by the discrete-time Fourier transform as
$$X(f) = DFT[x(k)]. \tag{5}$$
2. High-frequency components more than 10 kHz of $X(f)$ is averaged as
$$\mu = average[|X(f)|] \quad 10k < f < F_s/2 \tag{6}$$
where F_s is the sampling frequency.
3. Using the equivalence relation of
$$|N(f)| = \mu, \tag{7}$$
the spectral subtraction method is performed, and the output is downsampled by a decimator involving a low-pass filter. This downsampling is required to decrease the sampling rate to a commonly used one. The order of the operations of spectral subtraction and downsampling can be interchangeable.

4. Simulations

To verify the effectiveness of the proposed spectral subtraction method, we compare it with Boll's and Paliwal's methods. First, to visualize the fundamental properties of each method, we simulate each method for stationary noise conditions. And then we investigate the performance of each method for non-stationary

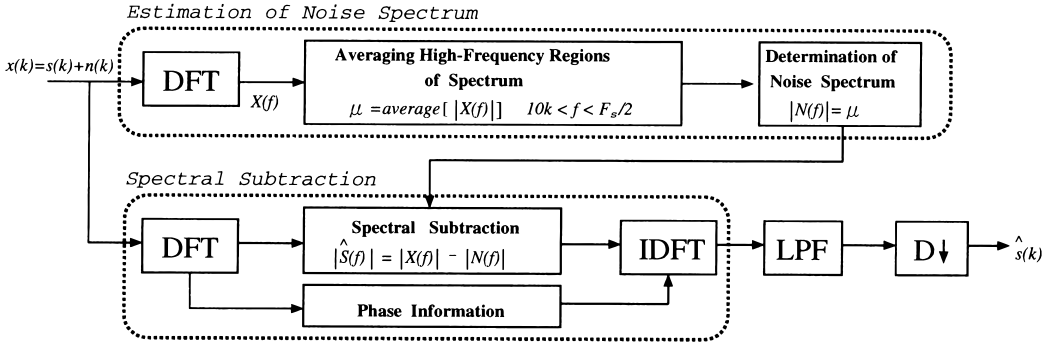

Fig. 2 A block diagram of the proposed spectral subtraction method (LPF: Low pass filter, D↓: Decimation).

noise conditions. Using MATLAB, Boll's method and Paliwal's method as well as the proposed method were programed by ourself.

4.1 Experimental Conditions

We prepared three kinds of speech data uttered by three male speakers. In this process, a sampling frequency of 30 kHz was used and each data length was commonly 6.7 s. These speech data were directly used for the proposed method with 50% overlapped segment. The each segment duration was 25.6 ms. By decimation with $D = 3$, however, the proposed method reconstructed a speech signal with a sampling frequency of 10 kHz. For Boll's and Paliwal's methods to be compared, the prepared speech data were downsampled to 10 kHz and used with the same segmenting way as in the proposed method.

We distinguished speech/non-speech segments in advance. Based on this, we evaluated the following segmental SNR.

[Improvement of Segmental SNR]

Generally, SNR is calculated through all data points as

$$SNR = 10\log_{10} \sum \frac{s(k)^2}{n(k)^2}. \tag{8}$$

The segmental SNR used in this work is a mean value of the SNRs evaluated in each speech segment. The improvement is given by subtracting the input SNR from the output SNR as

$$SNR_{improved} = SNR_{output} - SNR_{input}. \tag{9}$$

The result for each method was obtained by averaging those obtained on three kinds of speech data.

To assess the processed speech more validly, we further performed the following listening test.

[Listening Test]

In this test, the quality of speech was scored based on five levels as follows.

- 5 $\cdots$ very good
- 4 $\cdots$ good
- 3 $\cdots$ normal
- 2 $\cdots$ bad
- 1 $\cdots$ very bad

The speech processed by each method was subjected to eight listeners having normal hearing ability. And all scores obtained were averaged (totally 24 scores consisting of 3×8 were averaged) for each speech data.

4.2 Stationary Noise Case

First, we generated a stationary white noise and prepared noisy speech data with a SNR of 5 (dB), 0 (dB) and -5 (dB) for each of the three kinds of speech data. Using these speech data, we performed Boll's method, Paliwal's method and the proposed method.

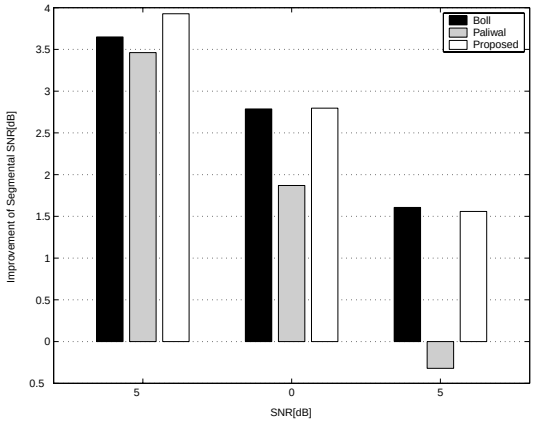

Fig. 3 Improvement of segmental SNR [dB] for stationary noise conditions.

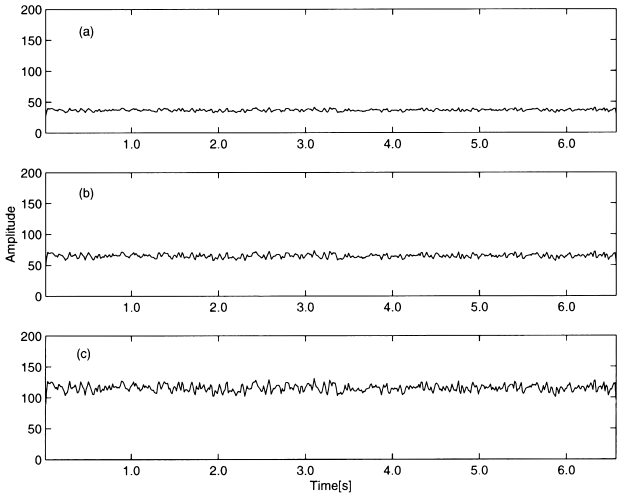

Fig. 4 The standard deviation of noise. (a) SNR=5 dB (b) SNR=0 dB (c) SNR=-5 dB

In Fig. 3, the results of the segmental SNRs evaluated have been summarized. For statistical investigation of hypothesis tests, it is observed from Fig. 3 that the performance of Boll's method is the same as that of the proposed method, while the performance of Paliwal's method is distinguished at a significance level of 1%. Paliwal's method deteriorates as SNR is increased. This may be because the high-order Yule-Walker equations required for the noise variance estimation in Paliwal's method have a tendency to make a singular correlation matrix leading to numerical unstability [5]. To confirm this, we investigated only the noise estimation process. Figure 4 shows the noise standard deviation evaluated from the noise included in each segment of the one kind of speech data (having three SNR levels). The noise standard deviation estimation by Paliwal's method and that by the proposed method are shown

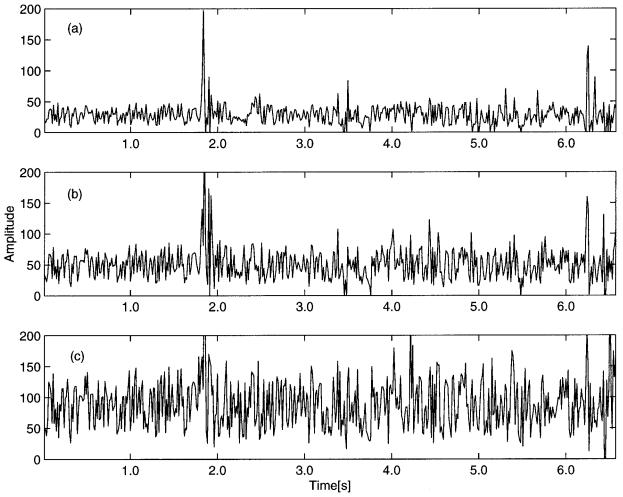

Fig. 5 Noise standard deviation estimated by Paliwal's method. (a) SNR=5 dB (b) SNR=0 dB (c) SNR=−5 dB

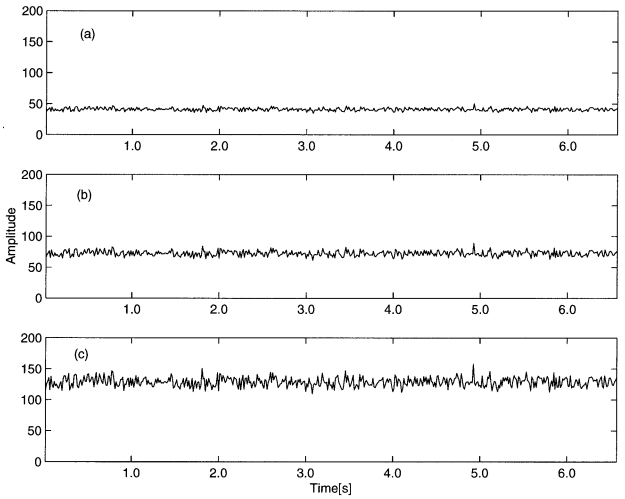

Fig. 6 Noise standard deviation estimated by the proposed method. (a) SNR=5 dB (b) SNR=0 dB (c) SNR=−5 dB

in Fig. 5 and Fig. 6, respectively. Obviously, the phenomenon of numerical unstability is observed in Fig. 5. Similar results for each method were obtained on the remaining two kinds of speech data.

Figure 7 has summarized the results of the listening tests performed. For statistical investigation of hypothesis tests again, it is observed from Fig. 7 that by processing the noisy speech data, Boll's and proposed methods provide an improvement, while Paliwal's method deteriorates in case of SNR=5 dB at a significance level of 1%. Figure 7 validates the results of Fig. 3 in case of SNR=5 dB. Figure 7, however, suggests that the performance of the proposed method is somewhat inferior to that of Boll's method, but is the same as that of Paliwal's method at lower SNRs.

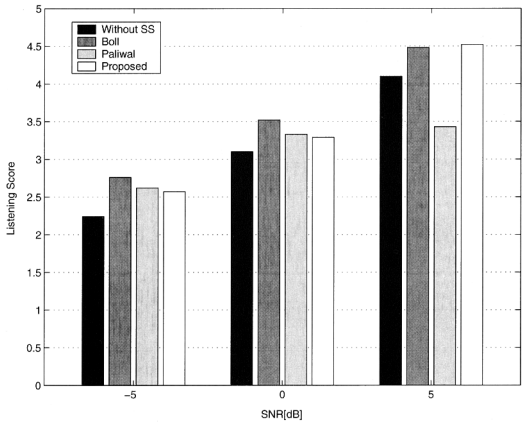

Fig. 7 Listening tests for stationary noise conditions. The SS denotes spectral subtraction.

4.3 Non-stationary Noise Case

We next simulated the three methods for non-stationary noise conditions. The speech data were corrupted by a white noise whose variance varies linearly with time. The non-stationary noise $n'(k)$ was generated based on

$$n'(k) = a(k)n(k) \quad 1 \le k \le L \tag{10}$$

$$a(k) = \begin{cases} \alpha_{High} \\ \quad +2(k-1)(\alpha_{Low} - \alpha_{High})/L \\ \qquad\qquad 1 \le k \le L/2 \\ \alpha_{Low} \\ \quad -2(k-L/2)(\alpha_{Low} - \alpha_{High})/L \\ \qquad\qquad L/2+1 \le k \le L \end{cases} \tag{11}$$

where L is the length of speech data generated and $n(k)$ is a stationary white noise. In Eq. (11), α_{High} and α_{Low} are a coefficient to be adjusted so that the SNR of the noisy speech data becomes a specified SNR. Three cases were considered. In the each case, the α_{High} and α_{Low} were set as follows;

- Case 1 $\cdots$ α_{High}(SNR=50 dB), α_{Low} (SNR= 0 dB)
- Case 2 $\cdots$ α_{High}(SNR=50 dB), α_{Low} (SNR= −5 dB)
- Case 3 $\cdots$ α_{High}(SNR=50 dB), α_{Low} (SNR= −10 dB)

where the value of SNR in each parenthesis corresponds to the specified SNR. If the difference between the SNR specified by α_{High} and that specified by α_{Low} becomes larger, the noise varies more rapidly (while the difference in Case 1 is 50 dB, the difference in Case 3 is 60 dB providing more rapid variation of the noise). The noise waveform in the above each case is shown in Fig. 8.

The results of the segmental SNRs evaluated have been summarized in Fig. 9. In the statistical sense at

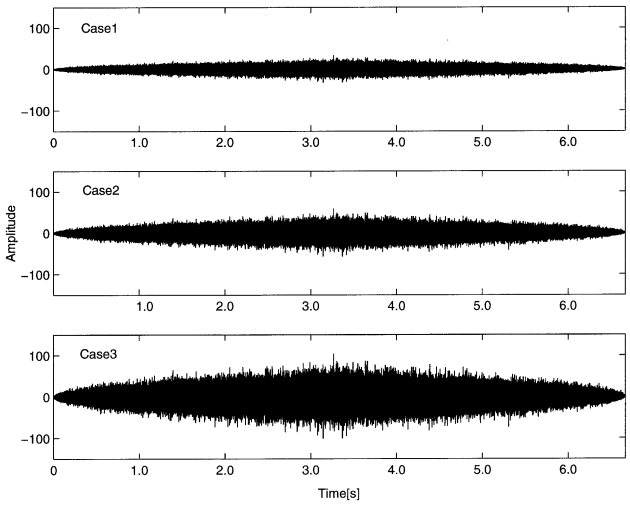

Fig. 8 Non-stationary noise in Cases 1-3.

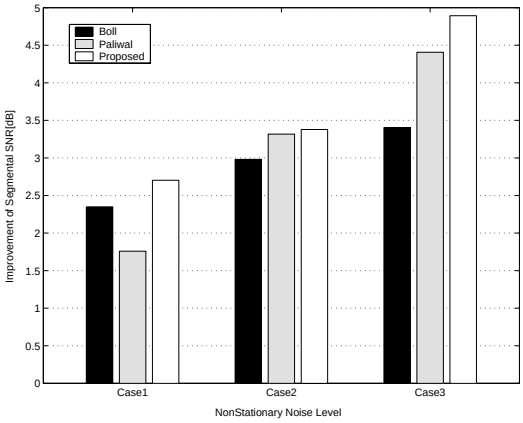

Fig. 9 Improvement of segmental SNR [dB] for non-stationary noise conditions.

a significance level of 1%, the following things are observed from Fig. 9. Compared with Boll's method, a performance improvement provided by the proposed method is emphasized in Case 3 rather than in Case 1. This is because in Case 1 the noise variance varies slowly. As the noise variance varies more rapidly as in Cases 2 and 3, Boll's method degrades the performance of noise suppression compared with the proposed method. This is an essential problem Boll's method possesses. As described in [4], Boll's method subtracts the noise components inadequately under non-stationary conditions. On the other hand, Paliwal's method becomes comparative with the proposed method as the noise varies more rapidly. This is because Paliwal's method gives an estimate of the noise variance in each analysis segment. However, the performance of Paliwal's method is also inferior to that of the proposed method.

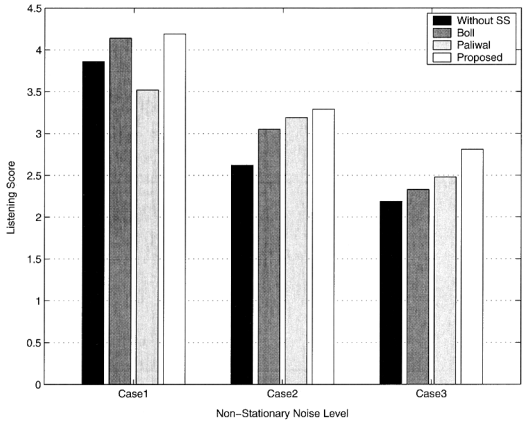

Fig. 10 Listening tests for non-stationary noise conditions. The SS denotes spectral subtraction.

The results of the listening tests are shown in Fig. 10. This figure also validates the results of Fig. 9.

5. Conclusion

In this paper we have proposed a spectral subtraction method based on the noise estimation using high-frequency regions of the noisy speech spectrum. With simple calculations the proposed method can invoke an adequate estimate of the noise included in the analysis segment, although it assumes that the noise is white. Experimental results have shown that the proposed method provides a performance improvement relative to Paliwal's method as well as Boll's method particularly for non-stationary noise conditions. Future work will aim to reduce the musical noise produced by the proposed method.

References

[1] J.S. Lim, ed., Speech Enhancement, Prentice Hall, Englewood Cliffs, 1983.
[2] J.S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," IEEE Trans. Acoust., Speech & Signal Process., vol.ASSP-26, no.5, Oct. 1978.
[3] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech & Signal Process., vol.ASSP-27, no.2, April 1979.
[4] K.K. Paliwal, "Estimation of noise variance for the noisy AR signal and its application in speech enhancement," IEEE Trans. Acoust., Speech & Signal Process., vol.36, no.2, February 1988.
[5] S. Kay, Modern Spectral Estimation: Theory and Application, Prentice Hall, Englewood Cliffs, 1988.