Single-Channel Speech Enhancement Based on Frequency Domain ALE

Isao Nakanishi
Faculty of Regional Sciences,
Tottori University
4-101 Koyama-minami, Tottori, 680-8551 Japan
Email: isao@rstu.jp

Yuudai Nagata, Yoshio Itoh, Yutaka Fukui
Faculty of Engineering,
Tottori University
4-101 Koyama-minami, Tottori, 680-8552 Japan
Email: {nagata@dacom2., itoh@, fukui@}ele.tottori-u.ac.jp

***Abstract*— In the present paper, a new single-channel speech enhancement system is proposed. The proposed system is based on frequency domain adaptive line enhancer; therefore, it is advantageous to non-stationary environments. Also, frequency domain decorrelation parameters are introduced and then adjusted independently. The performance of the proposed system is examined through computer simulations. The effectiveness of the proposed system is confirmed through computer simulations.**

I. Introduction

As voice communication systems are widely used in our daily life, speech enhancement techniques have attracted much attention. Among of them, single-channel speech enhancement is cost effective and suitable for miniaturizing the voice communication systems. The spectral subtraction (SS) method is well known as the single-channel speech enhancement technique [1] but it is weak in non-stationary environments. The reason is that the SS is a time-sharing method of one microphone: a noise spectrum is obtained in a speech pause and then it is subtracted from a noisy speech spectrum in the following speech existent period.

The authors have proposed another single-channel speech enhancement system which is based on frequency domain adaptive line enhancer (ALE) [2], [3]. The noise canceling using ALE had been already proposed in Ref.[4]. The authors introduced the frequency domain adaptive filter (FDAF) into the ALE in order to improve convergence speed of the ALE [5]. In the ALE, the input signal of the ADF is generated by delaying a desired signal. The time delay is constant and called decorrelation parameter, which makes the noise in the desired signal decorrelated with that in the input signal. On the other hand, the proposed frequency domain ALE enables setting the decorrelation parameter in frequency domain [6]. Resultingly, it reduces the computational complexity of the proposed system [7].

In the present paper, the authors propose to adjust the frequency domain decorrelation parameters according to the relation between the speech and the noise. The performance of the proposed system is examined through computer simulations.

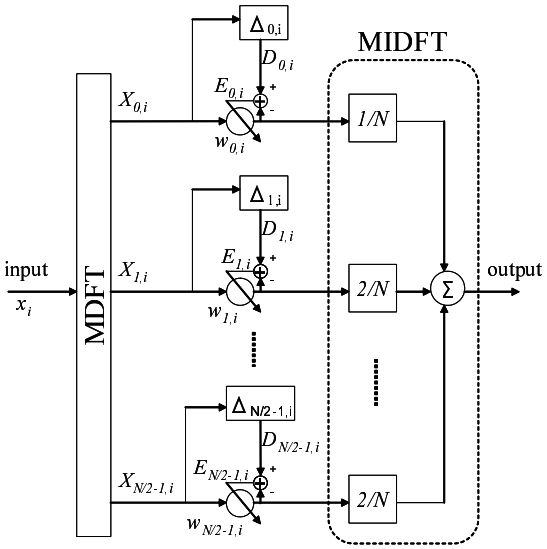


Fig. 1. The single-channel speech enhancement system based on frequency domain ALE.

II. Proposed Single-Channel Speech Enhancement System

The fundamental structure of the proposed single-channel speech enhancement system is based on the adaptive line enhancer (ALE). Such a structure is illustrated in Fig.1 where i and k are time and frequency indices.

The proposed system adopts modified DFT (MDFT) pair. The MDFT is obtained by simplifying the original DFT [9]. Moreover, for introducing the window function the MDFT is re-modified as

$$X_{k,i} = \sum_{n=0}^{N-1} x_{i-n} \ \cos(2\pi(n - N/2)k/N) \tag{1}$$

where $\cos(2\pi nk/N)$ in MDFT is shifted by $N/2$ [7]. N is the number of samples for DFT analysis and assumed to be even hereafter. Inverse MDFT (MIDFT) is defined as

$$x_i = \frac{X_{0,i}}{N} + \frac{2}{N} \sum_{k=1}^{N/2-1} X_{k,i} \ . \tag{2}$$

The MDFT pair requires only real-value operations and the MIDFT is achieved by summing the MDFT outputs, that is to say, the MDFT decomposes an input signal into harmonic signals while maintaining phase differences. Therefore, adaptive signal processing can be simply realized by adjusting the amplitude of the MDFT output signal.

By using the MDFT, the input signal x_i is decomposed into the harmonic signals $X_{k,i}$. The desired harmonic signals $D_{k,i}$ is obtained by delaying $X_{k,i}$ by $\Delta_{k,i}$ $(k = 0, 1, \cdots, N/2-1)$. They are frequency domain decorrelation parameters which can be set independently. These are examined in the next section.

The adaptive weight $w_{k,i}$ is multiplied by each $X_{k,i}$ and updated to reduce the error $E_{k,i}$ between $X_{k,i}$ and $D_{k,i}$:

$$E_{k,i} = D_{k,i} - w_{k,i} \cdot X_{k,i} \ . \tag{3}$$

For updating adaptive weight, a normalized step size algorithm is used because it is essential for achieving fast convergence in the FDAF [10]. Thus,

$$w_{k,i+1} = w_{k,i} + 2 \ \mu_{k,i} \cdot E_{k,i} \cdot X_{k,i} \ , \tag{4}$$

$$\mu_{k,i} = \frac{0.5}{|X_{k,i}|_p^2} \ , \tag{5}$$

where $\mu_{k,i}$ is a normalized step size, and $|X_{k,i}|_p$ is the maximum of each MDFT output.

Finally, adapted MDFT outputs are summed in the MIDFT and then a noise-reduced speech signal is reconstructed. Phase information on the input signal is also used in the output signal.

III. Optimal Setting of Frequency Domain Decorrelation Parameters

In the proposed system, the decorrelation parameters are inserted in the frequency domain and set independently. Therefore, it is important how to set them. Since the speech enhancement based on ALE is achieved by utilizing the difference of the correlation between the speech and the noise, the authors examine the frequency domain decorrelation parameters using the autocorrelation.

Not only the frequency domain signals of the speech but also those of the noise are periodic. Moreover, the period of the fundamental frequency signal becomes equal to N since the number of sampled data for MDFT is N. Therefore, it is natural that its autocorrelation has the maximum at the time lag of N. Similarly, the autocorrelations of kth harmonic signals have the maximum at the time lag of N/k. The time lag where the autocorrelation of a signal becomes maximum is effective for enhancing the signal. As a result, the optimal setting of the frequency domain decorrelation parameter for speech enhancement becomes

$$\Delta_k = \frac{N}{k} \ . \tag{6}$$

On the other hand, each autocorrelation becomes zero at the time lag of $N/(k \times 4)$. If the time lag causes zero correlation between two signals, it is effective for their decorrelation. As a result, it is found that the optimal setting of the frequency domain decorrelation parameter for noise suppression is different from that for speech enhancement and so it is defined as

$$\Delta_k = \langle \frac{N}{k \times 4} \rangle \tag{7}$$

where $\langle \ \rangle$ expresses the processing to an integer.

However, the above setting is not always effective. In general, the spectrum of the speech is maldistributed in low frequency range. When the noise is wideband and the noise element is dominant than the speech element in frequency domain, enhancing such frequency domain signals results in increasing the noise rather than enhancing the speech. In order to cope with the trade-off problem, the frequency domain decorrelation parameter is set equal to the pitch of the speech when the speech element is not dominant. Let

$$\Delta_k = Pitch \tag{8}$$

where $Pitch$ expresses the pitch period. It has been confirmed that the autocorrelations of speech harmonics also have local maximum at the time lag which is equal to the pitch [7]. The frequency domain decorrelation parameter which is larger than N/k is expected to suppress the noise element compared with Eq.(6) while enhancing the frequency domain signal.

IV. Adjusting of Frequency Domain Decorrelation Parameters

As mentioned above, the effective setting of the frequency domain decorrelation parameter for noise suppression is different from that for speech enhancement; therefore, the frequency domain decorrelation parameter must be adjusted according to the existence of the speech and the dominancy of the speech elements in frequency domain.

Concretely, the frequency domain decorrelation parameter is set by Eq.(7) for suppressing the noise in speech pauses. In the case of speech harmonics during speech existent periods, the decorrelation parameter is set by Eq.(6) or Eq.(8) according to the dominancy of speech elements in frequency domain signals. For non-harmonics, the decorrelation parameter is set by using Eq.(7) for suppressing such signals as the noise.

In the present paper, the dominancy of speech elements is determined by dividing frequency domain into four regions. Figure 2 shows a model of the distribution of speech elements. In the case of Japanese vowels, it is well known that their

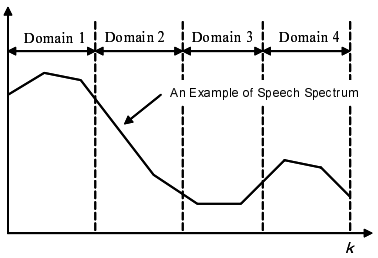


Fig. 2. A modeled distribution of speech elements.

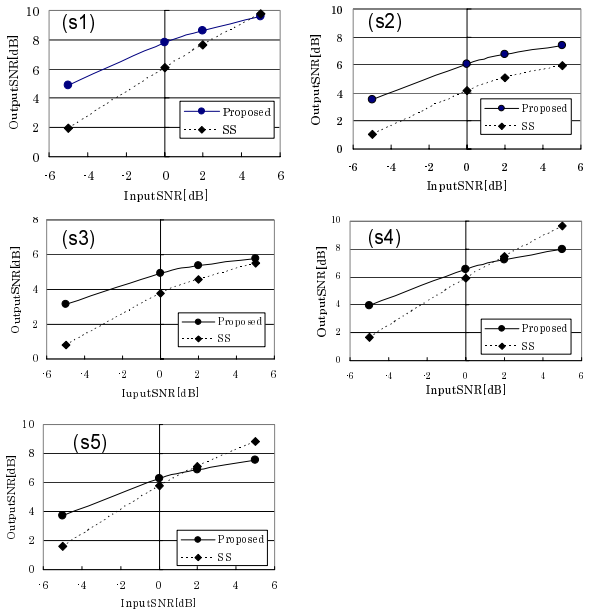


Fig. 3. Output SNR vs. Input SNR in a stationary case.

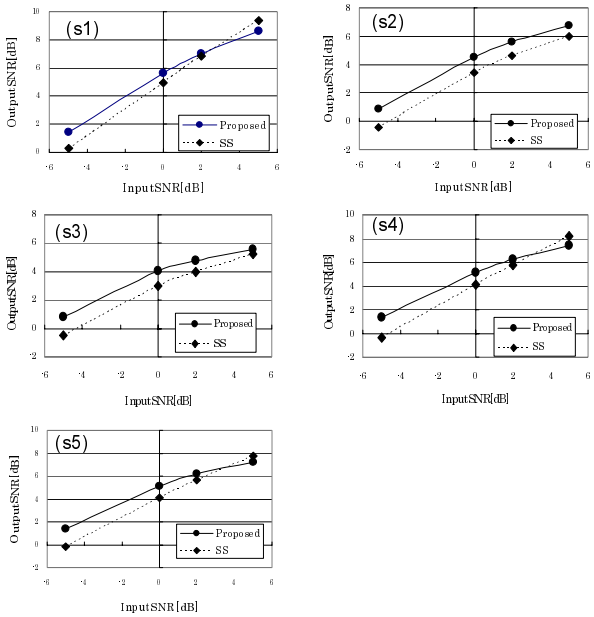


Fig. 4. Output SNR vs. Input SNR in a non-stationary case.

spectra have two mountains. Assuming that the noise is broad band, speech elements can be considered as dominant in Domain1 and inversely noise elements are dominant in Domain3. Domain2 and Domain4 correspond to their intermediate regions. As a result, harmonic signals in Domain1 are processed using Eq.(6) in order to enhance the speech elements. On the other hand, they are processed using Eq.(7) in Domain3 to suppress the noise elements. In Domain2 and Domain4, a compromise setting of speech enhancement and noise suppression is applied using Eq.(8).

In addition, the proposed system requires the detection of pitch and speech existence. There have been proposed several detection methods but simple detection methods are adopted in the present paper.

Firstly, the pitch is obtained by calculating the autocorrelation of the speech using the current and past 255 data: total 256 data sample by sample and then detecting the time lag where the autocorrelation becomes maximum.

Next, for detecting speech existence, the speech and noise level detectors proposed in Ref.[11] are adopted. The explanation of the detectors is omitted for lack of space. The principal speech elements can be assumed to be contained in the MDFT output at $k = 1$ (250.0 Hz) when $N = 32$ and 8 kHz sampling rate. When the output of the speech level detector is DS_i and that of the noise level one is DN_i, input SNR is estimated by

$$\text{Estimated input SNR} = 10 \log_{10} \frac{DS_i}{DN_i} \ . \qquad (9)$$

The estimated input SNR is compared with the threshold β. If it is larger than the threshold, then the speech is considered to be existent.

V. Simulations

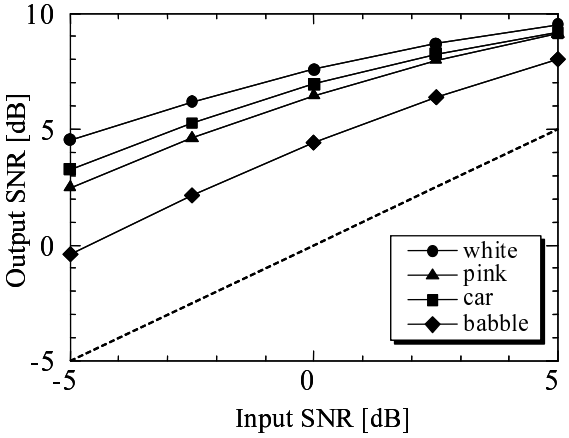
In order to verify the effectiveness of the proposed system, the authors carried out simulations. The following Japanese speech signals were used.

- s1:/watashiwasoreonozomu/
- s2:/soredejidainonagaregamienai/
- s3:/nazekoredakenokanegaugoitanoka/
- s4:/tanosekaidemosorewaonajidatoomoundesu/
- s5:/hyousyoudaidenoegaowatotemokireideshita/

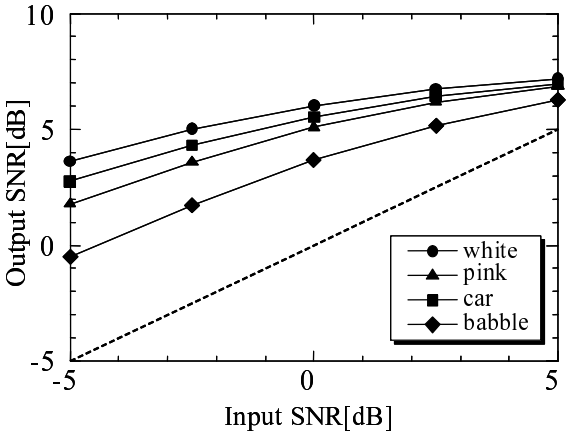
A white noise of variance 0.04^2 was used as a stationary noise and a noise generated by moving of a car was used as a non-stationary noise. It was assumed that Domain1 was $0 \leq k \leq 24$, Domain2 was $24 < k \leq 48$, Domain3 was $48 < k \leq 80$, and Domain4 was $80 < k$. The threshold for the detection of speech existence β was 12.5 dB. These were determined by trial and error in preliminary experiments. N was 256 and the Hamming window was used.

Results are summarized in Figs.3 and 4. For reference, the results by a modified spectral subtraction (SS) system [8] are also shown. The SS is well known as a speech noise reduction system in frequency domain and it requires only one microphone as well as the proposed system.

It is clear that the proposed system is more effective than the modified SS system when input SNR is less than about 2 dB. The SS system [1] is weak to reduce the non-stationary noise since a noise spectrum is preliminary estimated during a speech pause and then it is subtracted from the noise speech spectrum in the following speech existent period. The modified SS system is designed not to require such preliminary estimation but it assumes that the noise is relatively stationary compared with the speech [8]. In these simulations, such the assumption was not held, so that it brought about the



(a) Speech: s1



(b) Speech: s6

Fig. 5. Speech enhancement performance in various noise cases.

degradation of noise reduction performance. Also, since frequency domain signals is Domain3 were forcedly suppressed as the noise, it damages the original speech signal and so the performance of speech enhancement is degraded. It is remarkable in higher SNR conditions. The overall degradation of the performance of the proposed system in the non-stationary case is mainly due to the misdetection of speech existence.

Next, we evaluated the proposed system in the following various noises.

- White
- Pink
- Car interior
- Speech babble

These data were obtained from NOISEX database [12]. The speech signal s1 and s6:/sorekarajyuunenamari/ were used.

Results are shown in Fig.5. In the present paper, the dominancy of speech elements is determined based on an assumption that the noise is broad band. As a result, the speech enhancement performance was degraded when the noise was colored. In particular, such degradation of the performance was larger in the case of the speech babble. The speech babble was a speech-like noise generated by speaking of many people and so the characteristics of the noise was similar to those of the speech. Therefore, enhancing the speech resulted in enhancing the noise simultaneously.

VI. Conclusion

The single-channel speech enhancement system based on frequency domain ALE was proposed. Moreover, the optimal setting of frequency domain decorrelation parameters was proposed. They were adjusted according to the existence of the speech and the dominancy of the speech elements in frequency domain. The performance of the proposed system was evaluated using various speech and noise signals.

The detections of pitch and speech existence are important in the proposed system. More accurate and robust methods of detection must be further studied. To adjust the determination of the dominancy of speech elements depending on the SNR condition is a future work.

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Processing, vol.27, no.2, pp.113–120, Apr. 1979.

[2] I. Nakanishi, Y. Itoh, and Y. Fukui, "Accelerated frequency domain adaptive filter using modified DFT pair and its application to noise canceling," Proc. of the 2000 IEEE International Symposium on Circuits and Systems (ISCAS2000), Geneva Switzerland, vol.IV, pp.361–364, May 2000.

[3] I. Nakanishi, Y. Itoh, and Y. Fukui, "Noise reduction system based on frequency domain adaptive filter using modified DFT pair," Proc. of the 2001 IEEE International Symposium on Circuits and Systems (ISCAS2001), Sydney, Australia, vol.II, pp.737–740, May 2001.

[4] R. Sambur, "Adaptive noise canceling for speech signals," IEEE Trans. Acoust., Speech, Signal Processing, vol.26, no.5, pp.419–423, Oct. 1978.

[5] I. Nakanishi, Y. Hamahashi, Y. Itoh, and Y. Fukui, "A new structure of frequency domain adaptive filter with composite algorithm," IEICE Trans. Fundamental, vol.E81-A, no.4, pp.649–655, Apr. 1998.

[6] I. Nakanishi, T. Asakura, Y. Itoh and Y. Fukui, "Frequency domain de-correlation parameter in speech noise reduction system based on frequency domain adaptive line enhancer," Proc. of the 2004 47th IEEE Midwest Symposium on Circuits and Systems (MWSCAS2004), Hiroshima, Japan, vol.2, pp.13–16, July 2004.

[7] I. Nakanishi, T. Asakura, Y. Itoh and Y. Fukui, "Speech Noise Reduction System Based on Frequency Domain ALE Using Modified DFT Pair," Proc. of 2005 International Workshop on Acoustic Echo and Noise Control (IWAENC2005), Eindhoven, The Netherlands, pp.137-140, Sep. 2005.

[8] A. Kouda, T. Usagawa, and M. Ebata, "A new spectral subtraction method using the power change for noise spectrum estimation (in Japanese)," Journal of the Acoustical Society of Japan, vol.58, no.8, pp.493–500, Aug. 2002.

[9] S. Yoneda, I. Nakanishi, I. Sasaki, and A. Ogihara, "Switched-capacitor DFT and IDFT circuit," Int. J. Electronics, vol.67, no.6, pp.839–851, Dec. 1989.

[10] D.F. Marshall, W.K. Jenkins, and J.J. Murphy, "The use of orthogonal transforms for improving performance of adaptive filters," IEEE Trans. Circuits & Syst., vol.36, no.4, pp.474–484, Apr. 1989.

[11] I. Nakanishi, Y. Itoh, Y. Fukui and K. Fujii, "Noise reduction system using modified DFT pair," Proc. of the 2001 IEEE International Symposium on Circuits and Systems (ISCAS2001), Sydney, Australia, vol.II, pp.9–12, May 2001.

[12] http://spib.rice.edu/spib/select_noise.html