Speech Enhancement Based on Frequency Domain ALE with Adaptive De-Correlation Parameters

Isao Nakanishi, Hironori Namba, and Shigang Li

***Abstract*—We have proposed the speech enhancement method using the frequency domain adaptive line enhancer, which enables to independently set de-correlation parameters at frequency bins. In the conventional method, we divided the frequency band into four domains and set a de-correlation parameter at each domain. However, the spectral distribution of voices varies depending on gender and/or age differences. In this paper, we propose to adaptively control the de-correlation parameter according to the signal-to-noise ratio at each domain. Moreover, we propose to adjust the frequency width of each domain to a detected pitch (fundamental frequency). Their effectiveness is evaluated in simulations.**

***Index Terms*—Adaptive line enhancer, de-correlation parameter, frequency domain, speech enhancement.**

I. Introduction

Nowadays, mobile and smart phones are necessary in our daily life. They are used not only in doors but also out of doors, that is, in noisy environments. Thus, noise reduction or speech enhancement techniques are studied actively. Their aim is to reduce noise elements from a noisy speech signal and to enhance only speech elements.

As a typical method, Spectral Subtraction (SS) is well known [1]. In this method, the spectrum of a noise is estimated during a speech pause and then it is subtracted from the spectrum of a noisy speech for speech-existent periods. The SS method is efficient for miniaturization or cost reduction of portable devices since it needs only one microphone. On the other hand, the SS method needs preliminary estimation of a noise spectrum; therefore, it is difficult to apply the SS method in non-stationary environments.

For realizing speech enhancement using one microphone, another approach based on the adaptive line enhancer (ALE) has been proposed [2]. The ALE is one of applications of the adaptive digital filter (ADF) and its aim is to enhance sinusoidal waves buried in a broadband noise. This approach needs no preliminary estimation; therefore, it enables sequential processing and can be applied in non-stationary environments.

However, it is well known that the convergence speed of adaptive weights in the ADF is degraded when input signals are colored [3]. We have studied the frequency domain ADF (FDADF) in order to improve the convergence speed [4] and proposed a speech enhancement system based on the ALE using the FDADF [5].

In the ALE, a desired signal for the ADF is generated by delaying an input signal, and the delay time is defined as de-correlation parameter, which reduces the correlativity between noise elements in the desired signal and those in the input signal. As a result, only speech elements which are strongly correlated in both signals are extracted.

In general, the ALE has only one de-correlation parameter. On the other hand, in the proposed ALE using the FDADF (frequency domain ALE: FDALE), a desired signal and an input one are respectively decomposed into frequency signals, and each frequency signal is independently processed; therefore, it is possible to set a de-correlation parameter for each frequency signal. In other words, each frequency signal has an independent de-correlation parameter.

In the conventional system, based on the knowledge that there are four discriminative regions in the frequency band of human voices, the frequency band was divided into four domains, and de-correlation parameters were set according to the characteristic of each domain [5].

However, the spectral distribution of voices varies depending on gender and/or age differences. Thus, it is expected to bring better performance in speech enhancement by adaptively setting de-correlation parameters according to input signals. In this paper, we propose four methods of the adaptive setting of the de-correlation parameters and evaluate their effectiveness in simulations.

II. Speech Enhancement Based on FDALE

A. System Structure

Fig. 1 shows the structure of the proposed speech enhancement system based on FDALE. As transformation from a time-domain signal to frequency-domain signals, we use the modified DFT (MDFT) pair, which is a simplified version of the DFT pair [6]. The MDFT and its inverse transform (MIDFT) are defined as

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

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

where N is the number of sampled data analyzed in the DFT and is assumed to be even. n and i are time indices, and k is a frequency index.

Manuscript received September 4, 2012; revised November 12, 2012.

The authors are with the Graduate School of Engineering, Tottori, 680-8552 Japan (e-mail: nakanishi@ele.tottori-u.ac.jp, M09T3023@faraday.ele.tottori-u.ac.jp, li@ele.tottori-u.ac.jp).

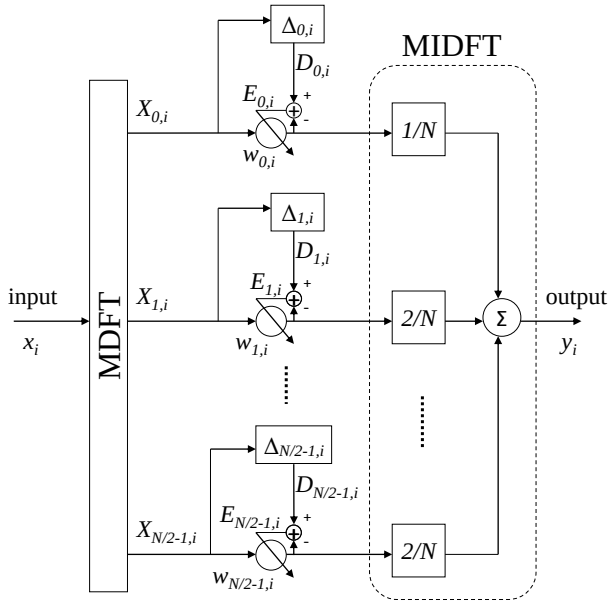


Fig. 1. Speech enhancement system based on FDALE.

In the case of using the window function, the MDFT is modified to

$$X_{k,i} = \sum_{n'=-N/2}^{N/2-1} x_{i-(n'+N/2)} \cos\left(2\pi n' k / N\right) \qquad (3)$$

where $n' = n - N/2$. For details, please refer to Ref. [5]. The modification brings the time delay of $N/2$ in the MDFT outputs.

In the proposed system, the amplitude of frequency signals is adaptively controlled by multiplying adaptive weights: $w_{k,i}$ to MDFT outputs: $X_{k,i}$ [4]. Therefore, the phase relation in the input signal: x_i is reused in the output signal: y_i in the same way as the SS method.

In general ALEs, the adaptive weights are updated by comparing a desired signal with an input signal; therefore, the weights are not independently updated and it results in slow convergence speed. In the FDADF, the adaptive weights are independently updated by comparing the spectra of the desired signal with those of the input one. This scheme is necessary to achieve faster convergence in the FDADF [4].

Concretely, the adaptive weight: $w_{k,i}$ at a frequency bin: k is updated so as to reduce an error spectrum: $E_{k,i}$ given by

$$E_{k,i} = D_{k,i} - w_{k,i} \cdot X_{k,i} \qquad (4)$$

where $D_{k,i}$, is a desired spectrum. The adaptive algorithm of the weights is defined as

$$w_{k,i+1} = w_{k,i} + \mu_{k,i} \cdot E_{k,i} \cdot X_{k,i} \qquad (5)$$

where $\mu_{k,i}$ is a step size that controls the convergence of the ADF. In this case, we use the following step size normalized by the power spectrum of the input signal.

$$\mu_{k,i} = \frac{1}{\left|X_{k,i}\right|_p^2} \qquad (6)$$

where $|X_{k,i}|_p$ is the maximum value for past N sampled data in the input spectrum. The normalization of the step size is also essential for achieving faster convergence in the FDADF [4].

$\Delta_{k,i}$ (k=0,1, ⋯;$N/2-1$) are de-correlation parameters in the frequency domain. These can be set independently; therefore, they enable optimal setting according to input signals.

B. Optimal Setting of Frequency Domain De-Correlation Parameters

General ALEs utilize the difference of correlativity between a speech signal and a noise one for enhancing the speech. On the other hand, in the proposed system, the noise signal is also decomposed into frequency signals as well as the speech signal, and thereby both frequency signals from the speech and the noise become sinusoidal; therefore, there is no correlativity difference between the frequency signal from the speech and that from the noise. Thus, we had examined optimal setting of frequency domain de-correlation parameters [5]. In the following, we summarize the points.

Let us consider the setting of frequency domain de-correlation parameters from the viewpoint of speech enhancement. It is natural that each frequency signal is periodical; therefore, its auto-correlation function becomes maximum at the time lag of integral multiple frequencies of N/k. As a result, to set the de-correlation parameters N/k brings enhancing the signals of frequency k.

$$\Delta_{k,i} = \frac{N}{k} \qquad (7)$$

where the value of the de-correlation parameter is limited to be integer; therefore, $\Delta_{k,i}$ is rounded off to an integer value. This setting is for enhancing speech elements.

On the other hand, for reducing frequency signals from a noise in the ALE, their correlativity should be reduced. Thus, the de-correlation parameters are set to be equivalent with the time lag of $N/(k \times 4)$, where the self-correlation function of the frequency signals certainly becomes zero.

$$\Delta_{k,i} = \left\langle \frac{N}{k \times 4} \right\rangle \qquad (8)$$

where if $N/(k \times 4)$ is not integer, it is multiplied by some integer value to obtain the integer value of the de-correlation parameter. Such processing is described as $\langle\rangle$. This setting is for reducing noise elements.

The above two settings are efficient when either a speech element or a noise element is dominant in a frequency signal. However, if both speech and noise elements are equal amount, to set the de-correlation parameter for enhancing the speech element simultaneously increases the noise element and vice versa. When both the elements are equivalently contained in a frequency signal, there exists a mutually contradictory problem: the signal is processed to enhance the speech element and simultaneously processed for reducing the noise element.

To cope with such a trade-off problem, the de-correlation parameter is set to the pitch (fundamental period) of an input signal.

$$\Delta_{k,i} = \mathrm{pitch} \qquad (9)$$

The auto-correlation function of frequency signals

becomes locally maximal at the time lag of not only N/k but also the pitch. Therefore, the de-correlation parameter of the pitch is also effective for enhancing the speech element. On the other hand, the pitch is generally larger than N/k. The correlativity of signals is decreased as the time lag between them is increased. Thus, a larger de-correlation parameter than N/k decreases the correlativity, that is, reduces the noise element while enhancing the speech element. This is a trade-off setting between enhancing speech elements and reducing noise ones.

C. Detection of Pitch and Speech Dominancy

As mentioned above, for applying the proposed settings of the de-correlation parameter, the detection of the pitch is needed. Speech signals are assumed to be stationary for a short duration: 20~40 ms. It corresponds to 256 samples when the sampling rate is 8 kHz. Therefore, by using the window of 256 samples, an auto-correlation function is sequentially calculated, and then the fundamental period of an input signal, that is, the pitch is detected. In addition, in order to avoid the miss-detection of the pitch, a restricted range is set in the detected value of the pitch. Concretely, assuming the pitch frequency is 84.5~369.7 Hz, the restricted range is set 2.7~11.8 ms. If the detected pitch value is out of this range, the previously detected one is used.

Furthermore, to know whether speech elements dominantly exist in a frequency signal is important for the proposed setting. The dominancy of the speech elements is detected by using the signal and noise level detectors that we had proposed. Please refer to Ref. [7] about the details of the detectors.

Assuming that the fundamental frequency of voices is from 84.5 Hz to 369.7 Hz, the output of the MDFT, that is, the frequency signal at k=1 (250 Hz) contains the principal elements of the voice when N=32 and 8 kHz sampling rate. By processing a frequency signal through the signal and noise level detectors, a signal-to-noise ratio (SNR) is sequentially estimated. If the estimated SNR is greater than a threshold, speech elements are assumed to be dominant in the frequency signal. The threshold value is empirically determined.

III. Adaptive Control of Frequency Domain De-Correlation Parameter

As mentioned in the previous section, there are three choices for setting the de-correlation parameter. It is important which is selected.

A. Method 1

In Ref. [5], based on the fact that there are four discriminative regions in the frequency band of human voices, we divided the whole frequency band into four domains, and set a de-correlation parameter according to the spectral characteristic of voices at each domain.

Assuming the sampling frequency is 8 kHz, we set the regions as follows.

- Domain 1 (D1) : 0 ~ 750 Hz (k : 0~24)
- Domain 2 (D2) : 750 ~ 1500 Hz (k : 25~48)
- Domain 3 (D3) : 1500 ~ 2500 Hz (k : 49~80)
- Domain 4 (D4) : 2500 ~ 4000 Hz (k : 81~127)

Speech elements are generally dominant in lower frequency bands; therefore, de-correlation parameters in the D1 are set for enhancing speech elements. On the other hand, in higher frequency bands, noise elements become dominant, so that de-correlation parameters should be set for reducing noise elements in the D3. In the D2 and D4, both speech and nose elements assumed to be equivalently dominant; therefore, the trade-off setting is adopted.

Furthermore, during speech pauses, the de-correlation parameters of all frequency signals must be set for reducing noise elements. The absence of speeches is sequentially detected by the SNR using the signal and noise level detectors in a similar way with the detection of the dominancy of speech elements in a frequency signal.

On the other hand, if the existence of speeches is detected, a fundamental frequency (pitch) is estimated, and then only harmonics and their neighboring frequency signals are processed using the above-mentioned de-correlation parameters. Other frequency signals are processed for reducing noise elements.

The settings of de-correlation parameters in the proposed system are illustrated in Fig. 2. This is called Method 1 in convenience.

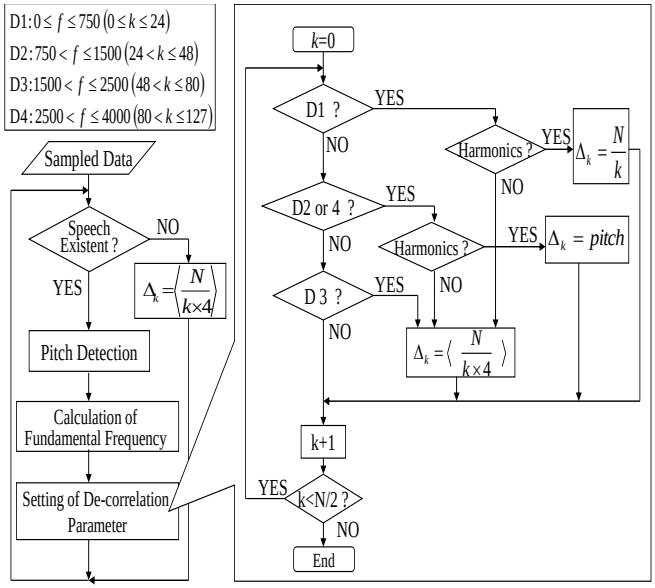


Fig. 2. Flow chart of Method 1.

B. Method 2

In Method 1, the setting of de-correlation parameters is fixed in each domain independently of the practical dominancy of speech elements. To control the setting according to the dominancy of speech elements could have effect on improving the performance of speech enhancement (noise reduction). Thus, we propose to switch two settings according to the SNR at each domain.

Concretely, a pair of speech and noise level detectors that is used for estimating the existence of speeches is also distributed to each domain, and then a SNR is estimated at each domain. In D1 and D2, that is, lower frequency band, if the estimated SNR is greater than a threshold, speech elements are assumed to be dominant, and then the setting for enhancing speech elements is applied. In the reverse case, the trade-off setting is adopted. In D3 and D4, that is, higher frequency band, if the estimated SNR is lower than the

threshold, the setting for reducing noise elements is applied, and if not so, the trade-off setting is adopted. This method is called Method 2, and its flow chart is shown in Fig. 3. The processing during speech pause and for harmonics is identical with that in Method 1.

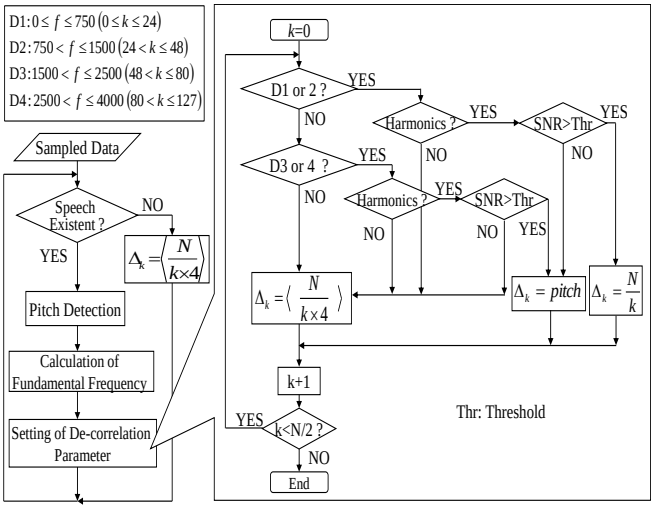


Fig. 3. Flow chart of Method 2.

C. Method 3

In Method 1 and 2, the frequency width of each domain is fixed. However, it is well known that pitch frequencies depend on gender and/or age differences. Therefore, the spectral characteristics of voices depend on individuals. To control the frequency width of each domain according to input signals could improve the speech enhancement performance.

Concretely, each width is defined by multiples of a detected pitch (fundamental) frequency. In this study, the starting frequencies of D2, D3, and D4 are set fivefold, twelvefold, and twenty-fivefold of the detected pitch frequency, respectively. On the other hand, the setting of a de-correlation parameter at each domain is fixed in the same way as of Method 1. This is called Method 3, of which flow chart is described in Fig. 4.

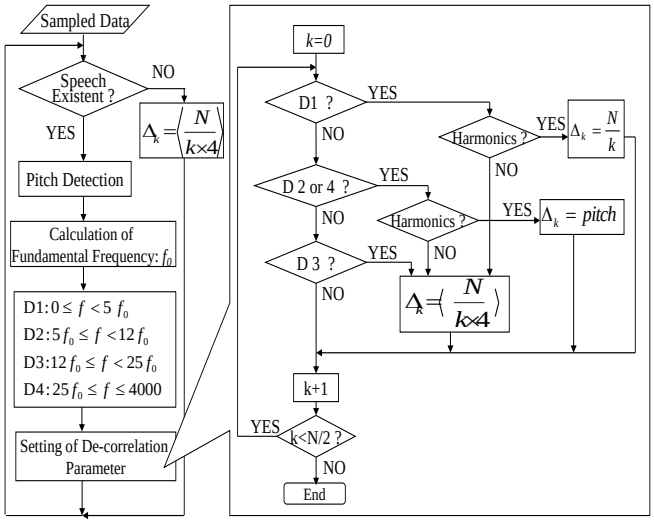


Fig. 4. Flow chart of Method 3.

D. Method 4

Finally, we propose to control both the setting of de-correlation parameters and frequency width in each domain. This method is just combining Method 2 and 3, and called Method 4.

In D1 and D2, the setting for enhancing speech elements and the trade-off setting are switched according to estimated SNRs, and in D3 and D4, the trade-off setting or the setting for reducing noise elements is chosen. In addition, the pitch (fundamental) frequency is extracted, and then the width of each domain is given by multiples of the detected fundamental frequency. In this study, the starting frequencies of D2, D3 and D4 are fivefold, nine fold, and fourteen fold of the fundamental frequency, respectively. Fig. 4 shows the flow chart.

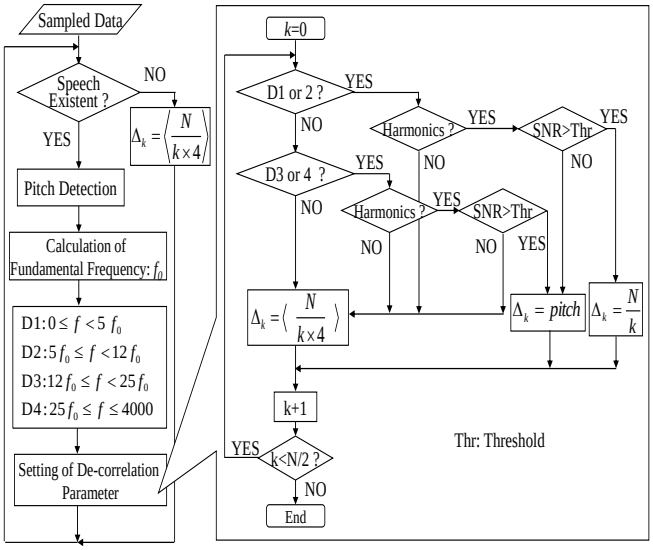


Fig. 5. Flow chart of Method 4.

The differences among the proposed four methods are summarized in Table I.

TABLE I: COMPARISON OF FOUR METHODS

		De-correlation Parameter	
		Fixed	Adaptive
Frequency Width	Fixed	Method 1	Method 2
	Adaptive	Method 3	Method 4

IV. SIMULATIONS

A. Conditions

In order to evaluate the proposed methods, we carried out simulations in speech enhancement. Three Japanese speeches (A, B, C) pronounced by three adult males and females were used as speech signal. An additional noise was a white one from Noisex-92 database [8], which was down-sampled from 19.98 kHz to 8 kHz. Sampling frequency was 8 kHz, quantized data length was 16 bit, and the DFT length N was 256. The SNR of all input signals was 0 dB. The threshold for determining the dominancy of speech elements were 12.5 dB.

B. Results

Table II shows improved SNRs (dB) in the case of male pronounce. These results are also compared in Fig. 6.

Next, improved SNRs (dB) in the case of female pronounce are shown in Table III. These are graphed in Fig. 7.

Finally, averaged values of the improved SNRs are shown in Table IV. These are compared in Fig. 8.

TABLE II: IMPROVED SNRS IN THE CASE OF MALE PRONOUNCE

	Method 1	Method 2	Method 3	Method 4
A	7.52	7.73	7.38	7.73
B	6.22	7.02	5.98	6.73
C	7.81	8.35	7.88	8.32
Average	7.18	7.70	7.08	7.59

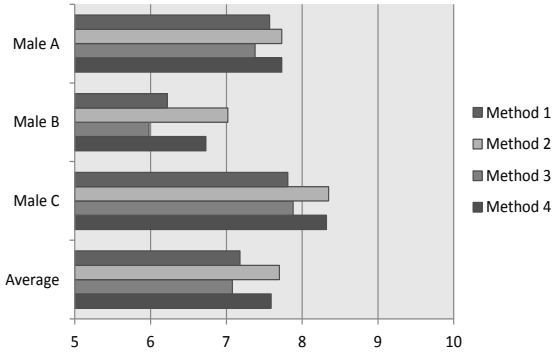


Fig. 6. Improved SNRs (dB) in male pronounce.

TABLE III: IMPROVED SNRS IN THE CASE OF FEMALE PRONOUNCE

	Method 1	Method 2	Method 3	Method 4
A	9.20	9.44	9.18	9.25
B	7.35	8.06	7.77	8.07
C	7.87	8.19	7.92	7.64
Average	8.14	8.56	8.29	8.31

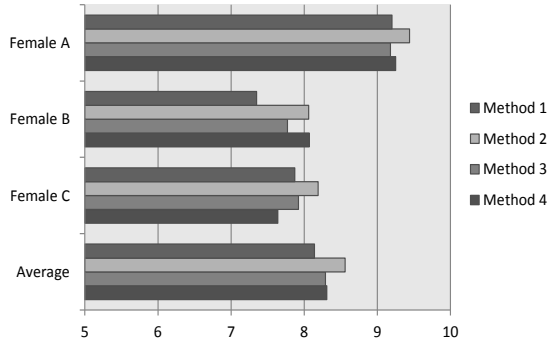


Fig. 7. Improved SNRs (dB) in female pronounce.

TABLE IV: AVERAGED VALUES OF IMPROVED SNRS

	Method 1	Method 2	Method 3	Method 4
Average	7.66	8.13	7.74	7.95

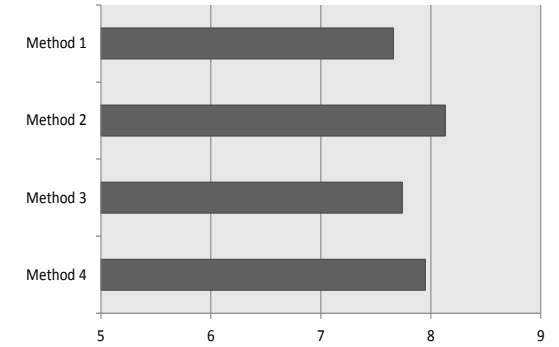


Fig. 8. Averaged improved SNRs (dB) in four methods.

C. Considerations

In this study, we expected that the method that controlled either the setting of the de-correlation parameter (Method 2) or the frequency width (Method 3) brought better performance, and the method that controlled both (Method 4) could bring the best performance. In the case of the speeches pronounced by female B, the results conformed to our intention but it was not achieved in other speeches.

From the results in the averaged values of improved SNRs, Method 2 achieved the best performance in almost all speeches. The result suggests that to control de-correlation parameters according to estimated SNRs brings better performance. However, the improvement is not sufficient. In this study, we used common thresholds for switching de-correlation parameters in all domains. In general, spectral elements are nonuniformly distributed; therefore, it may be effective to adjust the thresholds to the domains. It may be more efficient to choose the best one from three settings of the de-correlation parameter according to the estimated SNR in each domain while two settings were switched in this study.

Next, the improved SNR of Method 3 was sometimes inferior to Method 1. Therefore, we could not confirm the effect of controlling frequency widths according to a fundamental frequency. One of the reasons is low accuracy and/or stability of the pitch (fundamental frequency) detection. In order to confirm the effectiveness of Method 3, it is necessary to improve the accuracy of pitch detection. Moreover, excessive control of frequency widths might result in the degradation of performance.

Since Method 4 possessed the advantages of both Method 2 and 3, we expected that Method 4 was the best but it was almost inferior to Method 2. As mentioned above, the misdetection of pitch might degrade the performance. In addition, since the frequency width and the de-correlation parameter are simultaneously adjusted sample by sample in Method 4, such excessive adjusting might bring the degradation of performance. Furthermore, the bandwidth of the signal and noise level detectors used for estimating SNRs as well as in Method 2was fixed in all domains. To adjust the frequency width of each domain according to the detected fundamental frequency might cause the mismatch between the frequency width and the bandwidth of the detectors while such a mismatch is not caused in Method 2. It may be one of the reasons why Method 4 did not bring the best performance.

V. CONCLUSIONS

We had proposed a speech enhancement method using a frequency domain adaptive line enhancer (FDALE), which enabled to set de-correlation parameters independently at frequency bins. In the conventional method, the frequency band was divided into four domains and a suitable de-correlation parameter was set at each domain. However, in general, spectral distribution of voices varies depending on gender and/or age differences. It was required to cope with such a variation in order to improve speech enhancement performance.

In this paper, we proposed to control de-correlation parameters adaptively. Moreover, we proposed to adjust the

frequency width of each domain to a detected pitch frequency. Their effectiveness was evaluated in simulations.

Resultingly, the effect of controlling de-correlation parameters was confirmed but the effect of adjusting frequency widths was not fully recognized. The causes are as follows: low accuracy of detection of the pitch, excessive controlling of both the de-correlation parameter and the frequency width, and the mismatch between the frequency width and the bandwidth of signal and noise level detectors used for estimating SNRs.

To solve the problems is our future work. In addition, it will be a problem to divide the whole frequency band into more than four domains. Ultimately, the concept leads to a method in which the best setting is chosen from three ones; enhancing speech elements, reducing noise elements, and the trade-off setting at each frequency bin.

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

[3] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, N.J., 1985.

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

[5] I. Nakanishi, Y. Nagata, T. Asakura, Y. Itoh, and Y. Fukui, "Speech noise reduction system based on frequency domain ALE using windowed modified DFT pair," *IEICE Trans. Fundamentals*, vol. E89-A, no. 4, pp. 950-959, Apr. 2006.

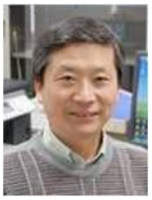
[6] S. Yoneda, I. Nakanishi, I. Sasaki, and A. Ogihara, "Switched-capacitor DFT and IDFT circuit," *Int. J. Electronics*, vol. 67, no. 6, pp. 839-851, Dec. 1989.

[7] I. Nakanishi, Y. Itoh, Y. Fukui, and K. Fujii, "Noise Reduction System Using Modified DFT Pair," in *Proc. of the 2001 IEEE International Symposium on Circuits and Systems (ISCAS2001)*, Sydney, Australia, vo. II, pp. 9-12, May 2001.

[8] NOISEX-92. [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

Isao Nakanishi was born in Osaka, Japan in 27 Dec. 1961. He received his B. E., M. E., and Dr. E. degrees in Electrical Engineering from Osaka Prefecture University, Japan in 1984, 1986, and 1997, respectively. He is now an associate professor in the Graduate School of Engineering, Tottori University, Japan. His research interests are in digital signal processing and biometrics. He is a member of the IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE) and the Information Processing Society of Japan (IPSJ).

Hironori Namba was in Okayama, Japan in 25 Oct. 1987. He received his B. E. and M. E. degrees in Electrical and Electronics Engineering from Tottori University, Japan in 2009 and 2012, respectively. His research interests are in speech signal processing.



Shigang Li was born in China in 3 Mar. 1963. He received his B. E. degree in Electrical Engineering from Beijing Tsinghua University, China in 1985. He received his M. E. and Dr. E. from Osaka University, Japan in 1990 and 1993, respectively. After receiving his Dr. E., he worked at Osaka University as a Research Associate. He became an associate professor with the Faculty of Information Sciences, Hiroshima City University, Japan in 1995. In 2001, he joined to the Faculty of Engineering, Iwate University, Japan. He has worked as a professor with the Graduate School of Engineering, Tottori University, Japan, since Oct. 2007. His research interests include computer/robot vision, intelligent transportation system and mixed reality systems.