Speech Noise Reduction Using Sequential Spectrum Detection Based on Modified DFT Pair

Isao Nakanishi Kensaku Fujii
Tottori University, Japan University of Hyogo, Japan
E-mail: nakanishi@ele.tottori-u.ac.jp fujiken@eng.u-hyogo.ac.jp

***Abstract*—A single-channel speech noise reduction using sequential spectrum detection based on a modified DFT pair is proposed in this paper. In the proposed method, a noise added speech is decomposed into frequency signals using the modified DFT and then signal and noise elements are sequentially detected from the decomposed frequency signal. Noise subtraction is operated by subtracting the noise element from the signal one. However, it needs a coefficient for compensating the difference between the detected noise level and the proper one. Therefore, an automatic setting of the coefficient is adopted in the noise subtraction. The effectiveness of the proposed noise reduction method is confirmed by subjective evaluations using processed signals.**

***Index Terms*—Speech noise reduction, single-channel, sequential spectrum detection, modified DFT pair**

I. Introduction

A lot of speech noise reduction methods have been proposed to improve the quality of speech communication in noisy environments. They are roughly categorized into two types. One is a one microphone (single-channel) type and the other is a multi-microphone type. As the multi-microphone type, an adaptive noise canceller and a microphone array are well known. They guarantee higher performance of noise reduction since they can refer to multiple input signals.

On the other hand, the single-channel type gains an advantage over miniaturization and cost reduction of portable digital equipments. Spectral Subtraction (SS) [1] is well known as a typical single-channel type and adopted in some of cellular phones. In this method, a noise spectrum is estimated during a speech pause and then it is subtracted from a noisy speech spectrum. However, strictly speaking, the SS is a quasi-single-channel type which time-shares one microphone for estimating both the noise spectrum and the noisy speech spectrum. Some modified methods have been proposed for coping with the problem [2], [3], [4].

We also have been proposed a new single-channel noise reduction system using sequential spectrum detection [5], [6]. A noisy speech signal is decomposed into frequency signals by using the modified DFT (MDFT) pair [7]. The MDFT pair is composed of FIR filters and is suitable for hardware implementation of sequential DFT operation comparing with the FFT while it is not suitable for applications with phase processing. Signal and noise spectra are estimated every sampled point by level detectors in each frequency signal. Noise reduction is sequentially operated by subtracting the detected noise spectrum from the detected signal spectrum [5]. However, a noise subtraction coefficient is required to compensate the difference between the detected noise spectrum and the proper one [6]. In this paper, we show a finding about an automatic setting of the noise subtraction coefficient. In addition, the effectiveness of the proposed system is examined by computer simulations of speech noise reduction and subjective evaluations.

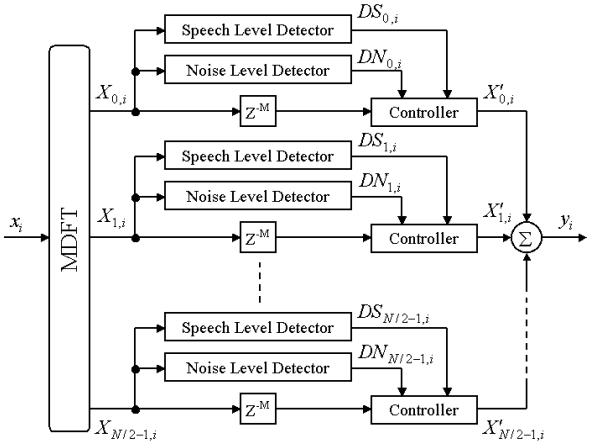

Fig. 1. System diagram.

II. Speech Noise Reduction Using Sequential Spectrum Detection

A. System Structure

A structure of the proposed noise reduction system is illustrated in Fig. 1. An input signal x_i is decomposed into frequency signals $X_{k,i}$ by using the MDFT. From each decomposed frequency signal, signal and noise spectra, $DS_{k,i}$, $DN_{k,i}$ are estimated using the level detectors every sampled point and then in the controller a speech spectrum is sequentially obtained by subtracting the estimated noise spectrum from the signal one. Noise reduced frequency signals $X'_{k,i}(k = 0, 1, \cdots N/2-1)$ are summed in the inverse MDFT (MIDFT) and as a result a noise reduced speech is obtained as y_i.

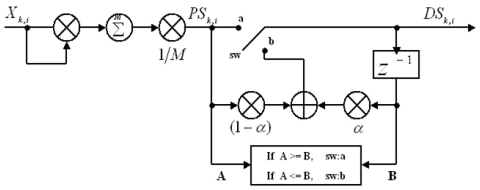

Fig. 2. Signal level detector.

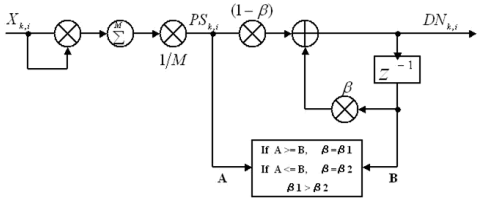

Fig. 3. Noise level detector.

B. MDFT Pair

The MDFT pair is derived from simplifying an original DFT pair [7]. The MDFT and MIDFT are respectively defined as

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

and

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

where the subscripts n and i are time indices and k is a frequency index. N is the number of sampled data in DFT analysis and it is assumed to be even in this paper.

The MDFT pair requires only real-value operations. Especially, the MDFT corresponds to the operations of a FIR filter and the MIDFT can be operated by summations. Therefore, the MDFT-MIDFT is realized by summing the outputs of the FIR filters. In other words, the MDFT decomposes the input signal into frequency signals keeping their phases; therefore, the MIDFT can be achieved by summing the MDFT outputs.

Moreover, it is well known that the spectrum side-lobe of a truncation function for the DFT analysis causes spectral leakages when the period of the truncation function is not equal to the fundamental period of the input signal. For reducing such a spectral leakage, it is also well known that a window function is effective. Introducing the window function into the MDFT had been proposed in Ref. [8]. The MDFT is rewritten as

$$X_{k,i} = \sum_{n=0}^{N-1} x_{i-n} \ \cos(2\pi(n - N/2)k/N) \ w(n) \tag{3}$$

where $w(n)$ is the window function. It is notable that the introduction of the window function yields the delay of $N/2$ samples [8].

C. Level Detectors

We assume that the noise spectrum is spread in wide frequency band and its rapidity of change is relatively slow. On the other hand, we also assume that the speech spectrum is located in lower frequency band and it varies rapidly as compared with the noise. We detect the speech and noise spectra using these assumptions.

1) Signal Level Detector: The diagram for detecting a signal level is shown in Fig. 2. The output of the MDFT is basically sinusoidal and its amplitude corresponds to the amplitude spectrum [7]. For full-wave rectifying, the output of the MDFT is squared, that is, the power spectrum is calculated. Additionally, the power spectrum is averaged with past $M-1$ data every sample for suppressing fluctuations.

$$PS_{k,i} = \frac{1}{M} \sum_{n=0}^{M-1} X_{k,i-n}^2 \tag{4}$$

The averaged power spectrum $PS_{k,i}$ is compared with the one-sample past output of the detector $DS_{k,i-1}$. If $PS_{k,i}$ is larger than $DS_{k,i-1}$, the switch (SW) is selected to "a" and the current output of the detector $DS_{k,i}$ is given by

$$DS_{k,i} = PS_{k,i}. \tag{5}$$

On the other hand, when $PS_{k,i}$ is smaller than $DS_{k,i-1}$, SW is switched to "b" and then the current output of the detector $DS_{k,i}$ is equal to the output of the following smoothing circuit with large time constant α.

$$DS_{k,i} = \alpha \ DS_{k,i-1} + (1-\alpha) \ PS_{k,i} \tag{6}$$

The signal level detector directly outputs rapid change when the amplitude of an input frequency signal increases and outputs slow change through the smoothing circuit when the amplitude of the input frequency signal decreases.

2) Noise Level Detector: The diagram of the noise level detector is shown in Fig. 3. The averaged power spectrum $PS_{k,i}$ is processed through the smoothing circuit defined as

$$DN_{k,i} = \beta \ DN_{k,i-1} + (1-\beta) \ PS_{k,i} \tag{7}$$

where the time constant β has different two values: β_1 and β_2 ($\beta_1 > \beta_2$).

When the averaged power spectrum $PS_{k,i}$ is larger than the one-sample past output of the detector $DN_{k,i-1}$, the larger time constant β_1 is selected. In the reverse case, β_2 is selected. To equalize the negative-going speed of the noise level detector with that of the signal level detector, β_2 is nearly equal with α.

Even if the averaged power spectrum rapidly changes, the output of the detector changes quite slowly. In general, the duration of each phoneme is relatively short; therefore, the output of the noise level detector keeps almost steady value without relation to the existence of the speech.

D. Spectrum Controller

Noise reduction is operated by subtracting the noise spectrum from the signal one in the spectrum controller (Controller). This concept is common to the SS method [1]. However, the proposed method repeats the noise subtraction every sample, so that it can cope with slow-changing noise.

The noise subtraction is defined as

$$SS_{k,i} = DS^{'}_{k,i} - \delta \ DN^{'}_{k,i} \tag{8}$$

where amplitude spectra $DS^{'}_{k,i}$ and $DN^{'}_{k,i}$ are used instead of the power spectra. Therefore, $SS_{k,i}$ corresponds to the amplitude spectrum of a noise reduced signal.

By the way, two detectors have different characteristics; therefore, it is not guaranteed that both detectors output the same level during speech pause. For compensating such a difference, a noise subtraction coefficient δ is adopted. If $SS_{k,i}$ becomes negative, it is treated as 0.

Next, the noise reduced spectrum $SS_{k,i}$ is returned to the frequency signal $X^{'}_{k,i}$ given by

$$X^{'}_{k,i} = \frac{SS_{k,i}}{DS^{'}_{k,i}} X_{k,i} \tag{9}$$

where the phase of the input $X_{k,i}$ is reused in the output. This is based on a phenomenon that auditory perception of human beings is relatively insensitive to the phase change.

In the MIDFT, all noise reduced frequency signals are summed and a noise reduced signal y_i is finally obtained.

$$y_i = \frac{X^{'}_{0,i}}{N} + \frac{2}{N} \sum_{k=1}^{N/2-1} X^{'}_{k,i} \tag{10}$$

E. Automatic Setting of δ

As mentioned above, the proposed method adopts the noise subtraction coefficient δ and its setting has a decisive influence on noise reduction performance. In this subsection, we propose an automatic setting of δ.

During speech pause, it is better to set δ larger in order not to cause residual noises. On the other hand, excessive subtraction damages to original speech elements in an input signal; therefore, smaller δ is preferable in speech existent periods. Consequently, δ must be adjusted according to existence or nonexistence of the speech.

This concept is realized by using the outputs of level detectors as follows:

$$\delta_{k,i} = \chi \ \frac{\sum_{n=0}^{m \times k} DN^{'}_{k,i-n}}{\sum_{n=0}^{m \times k} DS^{'}_{k,i-n}} \tag{11}$$

where χ is a coefficient for adjusting the value. The outputs of the detectors are accumulated and the number of samples $(m \times k)$ in the accumulation is proportional to the frequency index k since the fluctuation of the output level becomes larger as the frequency is higher.

If the speech existents, the denominator in Eq. (11) becomes large and then $\delta_{k,i}$ becomes small. On the other hand, the denominator becomes small in the speech pause and it results in large $\delta_{k,i}$.

New coefficients χ and m are set empirically but they are not highly dependent on kinds of speech and/or noise. This matter is confirmed in later experiments.

III. Subjective Evaluation

In order to confirm the effectiveness of the proposed method, we carried out subjective evaluations. Setting values of level detectors are summarized in Table I. In general,

TABLE I
SETTING VALUES OF LEVEL DETECTORS.

	α	β_1	β_2
$k < L$	0.991	0.9999999	0.998
$k \geq L$	0.998	0.99999	0.998

speech elements are almost located in lower frequency band; therefore, noise elements are dominant in higher frequency band. For this reason, the different values are set in lower $(k < l)$ and higher $(k \geq L)$ frequency bands. The boundary between the frequency bands is $L = 8$ in this paper. These values were found empirically [6].

We used two speeches pronounced by a male and a female in Multilingual Speech Database 2002 produced by NTT Advanced Technology Corporation. The content of the speech is the same. These are sampled at 8 kHz and digitized by 16 bit.

A white noise, a pink noise, a babble noise and a factory noise in NOISEX-92 were used, and the former two noises are of stationary and the latter two noises are of non-stationary.

The number of sampled data in the DFT analysis was $N = 128$, the number of sampled data for averaging the power spectrum was $M = 10$. χ and m for Eq. (11) were 100 and 7, respectively.

A. Evaluation Method

We used the noise reduced signals obtained by the proposed method in computer simulations. The number of subjects was 11, who were male students of our laboratory. They were requested to listen to the signals with a headphone in our laboratory room. First, the original speech was presented and then two different signals (A and B) were repeatedly presented twice. The subjects did not know which is A or B. Additionally, the order of presenting two signals was randomly shuffled. Standards for evaluation are as follows:

TABLE II
STANDARDS FOR EVALUATION.

+2	B is better than A
+1	B is relatively better than A
0	A is equal to B
-1	B is relatively worse than A
-2	B is worse than A

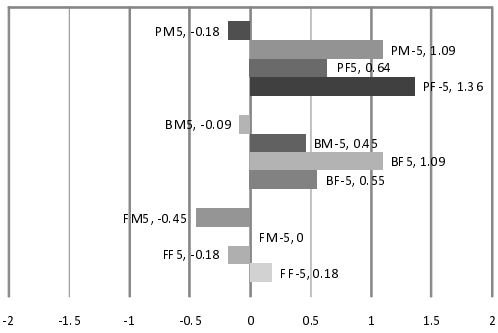

Fig. 4. Absolute evaluation of noise reduction performance.

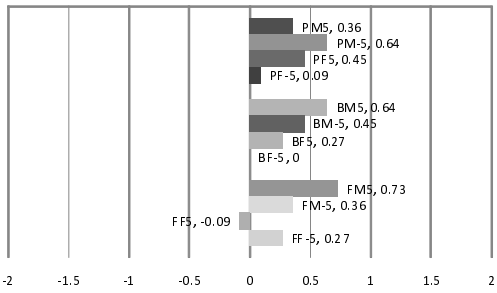

Fig. 5. Comparative evaluation comparing with the modified SS method.

B. Absolute Evaluation

To evaluate the noise reduction performance absolutely, we carried out the subjective evaluation using noise added signals and noise reduced signals.

The scores in evaluations are shown in Fig. 4. where three characters on the graph represent the type of the noise (Pink, Babble, or Factory), Male or Female, and SNR (5 dB or -5 dB). For instance, PM5 means the condition of the Pink noise, Male speaker and 5 dB.

In the case of the pink and babble noises, evaluation scores were more than 0; therefore, the effect of the noise reduction is confirmed. On the other hand, it was not evaluated in the case of the factory noise. The reason is that an artificial sound was generated by the residual noise in the processed signal by the proposed method and it led to such a bad evaluation.

The similar matter is observed in the case of SNR=5 dB where the evaluation scores were worse than those in other SNR conditions. The human being is able to hear the content of the noise added speech by using his/her auditory function in higher SNR conditions. Therefore, the newly-generated sound might be harsh for the subjects.

C. Relative Evaluation

Next, we compared the processed signals by the proposed method with those by the modified SS method [4] in the subjective evaluation. The modified SS method requires no preliminary detection of noise spectra; therefore, it is suitable for the comparative target of the proposed method. Assuming that an input signal consists of a stationary noise and a non-stationary speech, the small variation of the input signal power suggests a speech pause. Therefore, the noise power spectrum is estimated from the minimum local variance of spectral powers.

The results are shown in Fig. 5. The scores are small but almost positive; therefore, the noise reduced speeches by the proposed method is better than those by the modified SS method.

It is well known that the SS method also produces an artificial sound called the musical noise. In addition, based on our auditory evaluation, undulated or husky sounds were relatively perceived in noise reduced speeches by the modified SS method. These points might lead to worse rating.

IV. Conclusions

A new fully single-channel noise reduction method has been proposed. The noisy speech was decomposed into frequency signals by using the MDFT and signal and noise spectra were estimated by level detectors in each frequency signal. Noise reduction was sequentially achieved by subtracting the detected noise spectrum from the detected signal spectrum. The proposed method required a coefficient in the noise subtraction; therefore, automatic setting of the coefficient was also proposed. The effectiveness of the proposed method were confirmed by subjective evaluations.

It remains an issue that the proposed method produces artificial sounds in the noise reduced speech. Any countermeasure for suppressing such an artificial sound was not adopted in this paper. Partly because of the independent processing at each frequency signal, introducing the coordinated processing between frequency signals into the proposed method might suppress the residual noise and the excessive subtraction and then increase evaluated values.

References

[1] S. F. Boll, "Suppression of Acoustic Noise in Using Spectral Subtraction, " IEEE Trans. Acoust. , Speech, Signal Processing, vol. ASSP-27, no. 2, pp. 113-120, Apr. 1979.

[2] V. I. Djigan, P. Sovka, and R. Cmejla, "Modified Spectral Subtraction Based Speech Enhancement, " Proc. of 1999 International Workshop on Acoustic Echo and Noise Control (IWAENC1999), Pennsylvania, USA, pp. 64-67, 1999.

[3] S. Yoon and C. D. Yoo, "Speech Enhancement Based on Speech / Noise-Dominant, " IEICE Trans. Inf. & Syst. , vol. E85-D, no. 4, pp. 744-750, Apr. 2002.

[4] 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.

[5] I. Nakanishi, Y. Itoh, Y. Fukui, K. Fujii, "Noise Reduction System using Modified DFT pair, " Proc. of 2001 International Symposium on Circuits and System (ISCAS2001), Sydney, Australia, vol. II, pp. 9-12, May 2001.

[6] Y. Nagata, I. Nakanishi, Y. Itoh and Y. Fukui, "Performance Improvement of Noise Reduction System Using Modified DFT Pair, " Proc. of 2006 International Technical Conference on Circuits/Systems, Computers and Communications, Chiang Mai, Thailand, vol. II, pp. 365-368, Jul. 2006.

[7] 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.

[8] 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.