SINGLE CHANNEL BLIND DEREVERBERATION BASED ON AUTO-CORRELATION FUNCTIONS OF FRAME-WISE TIME SEQUENCES OF FREQUENCY COMPONENTS

[1]*Kenko OHTA*, [1]*Masuzo YANAGIDA*

[1]etf1704@mail4.doshisha.ac.jp
[1]Doshisha University, Dept. of Informatics and Intelligent Processing
1-3, Tatara-Miyakodani, Kyotanabe, Kyoto, 610-0321, Japan

ABSTRACT

This paper proposes a new blind dereverberation method based on auto-correlation function of frame-wise time serious of each frequency component. This method can handle frequency characteristics of sound paths. To realize dereverberation, it is necessary to estimate the delay time and frequency characteristics of reflection rate though most conventional methods assume flat frequency characteristics. The proposed method estimates these parameters based on auto-correlation function of frame-wise time sequence of each frequency component or a sequence of a spectral component in frequency spectra obtained every frame-shift interval. Comparing with methods that assume flat frequency characteristics, the proposed method improves dereverberation performance. The proposed method raises average segmental SNR 3.4 dB (from -2.4 dB to 0.9 dB) and reduces reverberation time from 390 ms to 109 ms.

1. INTRODUCTION

Performance of automatic speech recognition has reached a practical level in case a close contact microphone is used in quiet environments. In practical situations, however, recognition rate falls miserably due to environmental noises, reflected waves and so forth. There are two approaches for improving the recognition rate in practical environments. One is signal manipulation on the input signals and the other is introducing adaptation techniques in the recognition process. The proposed method is an approach classified to the former.

Inverse filtering of source-microphone transfer functions is widely employed for suppressing the effects of reflected waves [1], but this method can't be adopted to cases where the transfer functions from the source to microphones are not obtained and to time variant cases. Several methods of spectral subtraction have been proposed to cope with these cases [2] [3]. However, most of them require measurement of transfer functions among sources and microphones.

Takiguchi et al. proposed an adaptive method which doesn't require transfer functions [4]. The method, however, can't yield sufficient improvement in case the reverberation time is long, though it employs acoustic models obtained onsite.

Nakatani et al. proposed a method based on harmonic structure[5]. Their method constitutes an inverse filter based on a large number of reverberant speech data. The method, however, takes a lot of time to design an accurate inverse filter. Hence, it is difficult to put their method into practical use.

There are some methods based on linear prediction analysis [6] [7]. These methods require obtaining linear prediction coefficients of the speech beforehand.

The proposed method, however, does not require transfer function, inverse filters nor linear prediction coefficients.

The rest of the current paper consists of the following four sections. Section 2 shows the principle of the proposed method. Section 3 shows performance of the proposed method on simulated and actual data. Section 4 discusses dereverberation scheme by controlling the number of frequency bins processed as a group. Section 5 concludes the paper.

2. DESCRIPTION OF THE PROPOSED METHOD

2.1. Basis of Dereverberation

A signal $x(t)$ received by a microphone generally consists of several waves from sources including the direct wave and reflected waves. The signal $x(t)$ from a source is represented by convolution of source signal $s(t)$ and impulse response $h(t)$ from the source to the microphone. The signal $x(t)$ is expressed as:

$$x(t) = s(t) * h(t) \tag{1}$$

where * denotes convolution. Taking short term Fourier transform, eq. (1) can be rewritten as follows:

$$X(\omega_n, k) = S(\omega_n, k)H(\omega_n, k) \tag{2}$$

where ω_n denotes n-th frequency bin and k denotes frame ID. Here, $H(\omega_n, k)$ can be divided into the direct path

component $D(\omega_n, k)$ and the total sum of reflection components $R(\omega_n, k)$. So, eq. (2) is rewritten as follows:

$$X(\omega_n, k) = S(\omega_n, k)\{D(\omega_n, k) + R(\omega_n, k)\} \quad (3)$$

The frequency spectrum $X(\omega_n, k)$ at frequency ω_n in frame k of the received signal can be approximated by the convolution of the frequency spectrum $S(\omega_n, k)$ with $\alpha_{\omega_n}(k)$, impulse response of the integrated propagation paths at frequency ω_n.

$$X(\omega_n, k) \cong \sum_{l=0}^{L_n} \alpha_{\omega_n}(l) S(\omega_n, k-l) \quad (4)$$

where, L_n represents the time delay of n-th frequency bin and $\alpha_{\omega_n}(l)$ denotes the decay by distance and reflection characteristics at frequency ω_n.Then, the component for $l = 0$ in the summation for $X(\omega_n, k)$ can be regarded as the direct path component $S(\omega_n)D(\omega_n)$ and the other components in $X(\omega_n, k)$ can be regarded as the total sum of reflection components $S(\omega_n)R(\omega_n)$. So, eq.(4) can be rewritten as follows:

$$X(\omega_n, k) \cong \alpha_{\omega_n}(0) S(\omega_n, k) + \sum_{l=1}^{L_n} \alpha_{\omega_n}(l) S(\omega_n, k-l) \quad (5)$$

Here, $\alpha_{\omega_n}(0)$ is regarded to be unity and eq.(5) is rewritten as follows:

$$\hat{S}(\omega_n, k) \cong X(\omega_n, k) - \sum_{l=1}^{L_n} \alpha_{\omega_n}(l) S(\omega_n, k-l) \quad (6)$$

As the result of this processing, the frequency spectrum of source signal is estimated.

2.2. Estimation of the Delay Time and Decay Rate

Auto-correlation function is employed for estimating the delay time and decay rate for each frequency bin. The frequency spectrum of a received signal can be approximated by eq. (4). So, the auto-correlation function of the received signal is expressed as:

$$\phi_{XX}(\omega_n, k) = \sum_{m=0}^{M} \left(\sum_{l_1=0}^{L_n} \alpha_{\omega_n}(l_1) S(\omega_n, k-l_1) \right) \left(\sum_{l_2=0}^{L_n} \alpha_{\omega_n}(l_2) S(\omega_n, k+m-l_2) \right) \quad (7)$$

Then, separating the production of the inner $\sum$ into two cases, one for $l_1 = l_2 - k$ and the other for $l_1 \neq l_2 - k$, the above-mentioned equation is rewritten as follows:

$$\phi_{XX}(\omega_n, k) = \sum_{m=0}^{M} \left(\sum_{l=0}^{L_n - k} \alpha_{\omega_n}(l) \alpha_{\omega_n}(l+k) S(\omega_n, m-l)^2 \right) + \sum_{m=0}^{M} \left(\sum_{\substack{l_1=0 \\ l_1 \neq l_2 - k}}^{L_n} \alpha_{\omega_n}(l_1) S(\omega_n, m-l_1) \right) \left(\sum_{\substack{l_2=0 \\ l_1 \neq l_2 - k}}^{L_n} \alpha_{\omega_n}(l_2) S(\omega_n, k+m-l_2) \right) \quad (8)$$

Here, if each component of $S(\cdot, m-l)$ is mutually independent or $S(\cdot, m-l_1)S(\cdot, k+m-l_2)$ is enough smaller than $S(\cdot, m-l)^2$, $\phi_{XX}(\omega_n, k)$ can be approximated as:

$$\phi_{XX}(\omega_n, k) \cong \sum_{m=0}^{M} \left\{ \left(\sum_{l=0}^{L_n - k} \alpha_{\omega_n}(l) \alpha_{\omega_n}(l+k) S(\omega_n, m-l)^2 \right) \right\} \quad (9)$$

Equation (9) is normalized by dividing by $\phi_{XX}(0)$, so the equation is shown as follows:

$$\Phi_{XX}(\omega_n, k) = \frac{\sum_{m=0}^{M} \left\{ \left(\sum_{l=0}^{L_n - k} \alpha_{\omega_n}(l) \alpha_{\omega_n}(l+k) S(\omega_n, m-l)^2 \right) \right\}}{\sum_{m=0}^{M} \left\{ \left(\sum_{l=0}^{L_n} \alpha_{\omega_n}(l) \alpha_{\omega_n}(l) S(\omega_n, m-l)^2 \right) \right\}} = f_n(\omega_n, k) \quad (10)$$

However, $\Phi_{XX}(\omega_n, k)$ is nearly zero for k larger than L_n. The effective reverberation time L_n for each frequency bin is assumed as the point beyond which $\Phi_{XX}(\omega_n, k)$ is smaller than ϵ, where ϵ is an enough small value. Moreover, $\Phi_{XX}(\omega_n, k)$ is regarded as the decay rate for delay k for frequency ω_n.

$$\alpha_{\omega_n}(k) = \Phi_{XX}(\omega_n, k) \qquad k = 1 \cdots L_n \quad (11)$$

2.3. Subtracting Reflected Waves on the Power Spectrum

There should be clean frames without effects of reflection at the beginning of utterance before the first reflection reaches the microphone. Dereverberation can be achieved by eq. (6). A processing scheme for eq. (6) is depicted in Fig. 1, where the abscissa corresponds to the time axis plotted frame by frame and the ordinate symbolically represents amplitude at frequency ω_n. Tick width over the time sequences of the amplitude component represents effective reverberation time L_n. The gray bar in the upper half of this figure represents the direct component and black bars represent the reflected wave components. The lower half of this figure shows the time sequences of the amplitude component after subtracting the reflected components of $S(\omega_n, 0)$.

3. PERFORMANCE EVALUATION

3.1. Evaluation Using Simulated Data

Simulation is performed to compare the proposed method with the method that assumes flat frequency characteristics at reflection on wall. Figure 2 shows spectral comparison of simulated data. Comparing (c) and (d) in Fig. 2, we can see that the proposed method well suppresses oversubtraction. As an index for evaluating the performance,

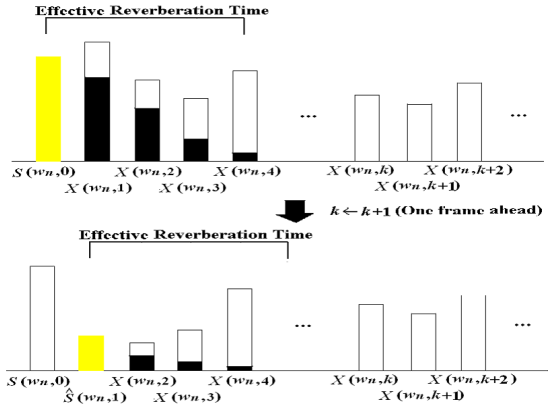

Figure 1: *Spectral subtraction on the time sequence of a frequency bin corresponding to a certain frequency* ω_n.

employed is segmental SNR defined as follows:

$$SNR(k)=\frac{\sum_{n=0}^{N-1}|S(\omega_n,k)|^2}{\sum_{n=0}^{N-1}|S(\omega_n,k)-\hat{S}(\omega_n,k)|^2} \quad (12)$$

where, N denotes the number of frequency bins on the output of FFT. The average segmental SNR of speech signal before processing is -2.4 dB. The average segmental SNR of the speech signal processed by the proposed method is 0.7 dB, while that by the conventional method assuming flat frequency characteristics is 0.5 dB. Figure 3 shows estimated frequency characteristics of reverberation time by the two methods. Here, short term Fourier transform is performed under the condition of 64 ms frame length and 4 ms shifting interval. From Fig. 3, we see that the longest reverberation time is about 800 ms around $3 \sim 3.5$ kHz. We can see that the frequency having the longest reverberation time in Fig. 3 corresponds to the frequency having long lasting tails in Fig. 2 b). If we take the dotted line given by the conventional method assuming flat frequency characteristics as the reverberation time, it is clear that over-subtraction occurs because the estimated reverberation time is too long.

3.2. Evaluation Using Actual Data

Evaluation data are recorded in a small cabin (230 cm $\times$ 380 cm, H :218 cm). An evaluation result for a Time Stretched Pulse (TSP) is shown in Fig. 4, which shows spectrograms of the TSP before and after processing. Figure 5 shows frequency characteristics of the reverberation time estimated by the proposed method. From this figure, the longest reverberation time is estimated as about 740 ms. Comparing Fig. 4 with 5, the contour of estimated frequency characteristics of the reverberation time is similar to the outline of spectrogram of the TSP. So, we may be able to suppose that frequency characteristics

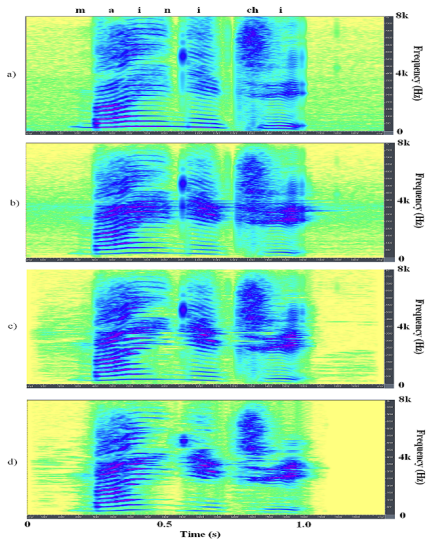

Figure 2: *Performance comparison on spectrograms for simulated data.(a: Original speech, b: reverberant speech, c: dereverberated speech by the proposed method, d: dereverberated speech by a conventional method assuming flat frequency characteristics)*

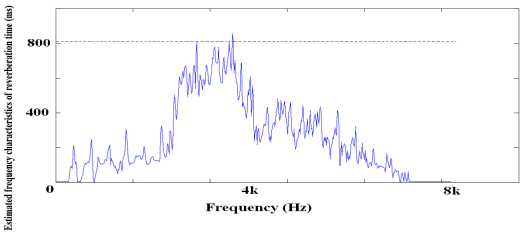

Figure 3: *Estimated frequency characteristics of reverberation time. (The solid line represents the results obtained by the proposed method and the horizontal dotted line is that obtained by the conventional method assuming flat frequency characteristics.)*

are successfully estimated in case of actual data. To confirm the performance of the proposed method, reverberation curve is depicted in Fig. 6, from which reverberation time appears to be 390 ms. On the other hand, reverberation time is shortened to be 109 ms using the proposed method. Figure 6 shows that the proposed method drastically reduces initial reflections.

4. DISCUSSION

In previous sections, the proposed dereverberation method is described to be processed each frequency bin. As the frequency characteristics are supposed to be similar within a narrow frequency range, the processing scheme expressed in eq. (6) can be handled in a group within narrow fre-

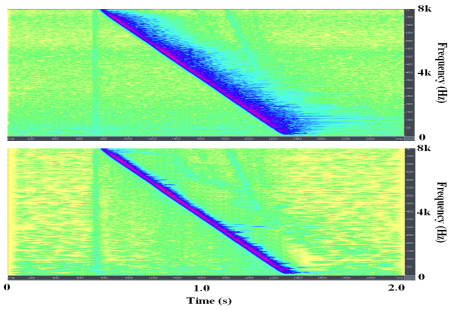

Figure 4: *Spectrogram of the TSP. (Top: Received, Bottom: Processed by the proposed method)*

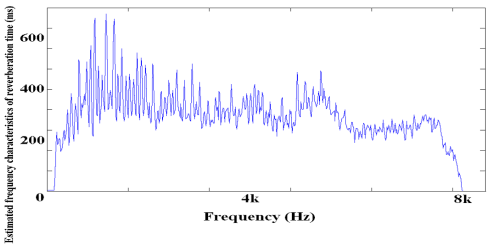

Figure 5: *Estimated frequency characteristics of the reverberation time.*

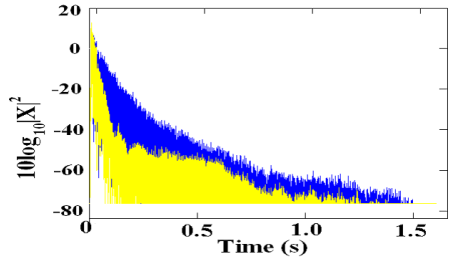

Figure 6: *Reverberation curves.(Black area: Received, Gray area: Processed by the proposed method)*

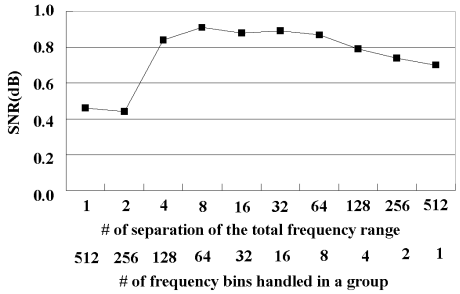

Figure 7: *Average segmental SNRs obtained for several separation numbers of the total frequency range.*

quency ranges. So, to get better dereverberation performance, we try a scheme in which some of adjacent frequency bins are by varying simultaneously processed. Figure 7 shows average segmental SNRs obtained for various the separation numbers for dividing frequency bins. From this figure, the best performance seems to be obtained by the proposed method at 4 to 64 separations in case we use FFT of 1024 points.

5. CONCLUSION

This paper proposes a single channel blind dereverberation method based on auto-correlation functions of time sequences of frequency components on running power spectra. The proposed method estimates the delay time and the frequency characteristics of decay rate. So, the proposed method achieves dereverberation without any premeasurement or *a priori* information. From the performance evaluation on simulated data and actual data, the proposed method yields better results than conventional spectral subtraction methods.

6. REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, pp. 145–152, 1988.

[2] A. Baba, D. Matsumoto, A. Lee, and K. Shikano, "Recognition of speech with dereverberation by spectrum subtruction in home environment," in *Proc. of Acoustic Society of Japan*, Okinawa, Sept. 2004, pp. 17–18 in Japanese.

[3] D. Matsumoto, A. Baba, A. Lee, H. Saruwatari, and K. Shikano, "Speech recognition of distant talking for human-robot speech interface," in *Proc. of Acoustic Society of Japan*, Okinawa, Sept. 2004, pp. 9–10 in Japanese.

[4] T. Takiguchi, S. Nakamura, Q. Huo, and K. Shikano, "Model adaptation based on HMM decomposition for reverberant speech recognition," in *Proc. of ICASSP97*, Munich, Apr. 1997, pp. 827–830.

[5] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. of ICASSP2003*, Hong Kong, Apr. 2003, pp. 92–95.

[6] K. Kinoshita, T. Nakatani, and M. Miyosi, "Single channel blind dereverberation using multi-step forward linear prediction," in *Proc. of Acoustic Society of Japan*, Tokyo, Mar. 2006, pp. 511–512 in Japanese.

[7] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. of ICASSP2001*, Salt Lake City, May 2001.