SPEECH DEREVERBERATION BY COMBINING MINT-BASED BLIND DECONVOLUTION AND MODIFIED SPECTRAL SUBTRACTION

Ken'ichi Furuya, Sumitaka Sakauchi, and Akitoshi Kataoka

NTT Cyber Space Laboratories, NTT Corporation
3-9-11, Midori-cho, Musashino-shi, Tokyo 180-8585, Japan

ABSTRACT

A dereverberation technique is developed to provide an alternative means of reducing reverberation in speech signals. The conventional MINT (the multiple-input/output inverse-filtering theorem) method uses the room impulse responses to calculate the inverse filters, so it cannot recover speech signals in practice, where the room impulse responses are unknown in advance. Our method blindly estimates the inverse filters by computing the correlation matrix between input signals that can be observed, instead of room impulse responses. We also combine the inverse filtering with modified spectral subtraction against the estimation error of inverse filters used in the field. The performance of the proposed method is demonstrated using actual room impulse responses.

1. INTRODUCTION

When a speaker is some distance away from the microphone in a teleconference, the speech signal is distorted by room reverberation, so it is less intelligible to listeners. One theoretical way to achieve nearly perfect dereverberation of speech is to perform inverse filtering using several microphones based on the multiple-input/output inverse-filtering theorem (MINT) [1]. The MINT method computes stable and accurate inverse filters of room impulse responses that may be in the nonminimum phase [2]. This method requires that room impulse responses of sound transmission channels are known in advance, but there has been no practical way to know the impulse responses between the speaker and the microphones.

A number of multichannel blind deconvolution methods, [3] – [7], without measuring room impulse responses have recently been developed for speech dereverberation. We introduced a MINT-based blind deconvolution method [8]. However, blind deconvolution methods based on inverse filters including the MINT-based method are generally not so robust against small errors in the estimation of inverse filters and can hardly improve on the tail part of reverberation in the actual world where impulse responses can be always fluctuating.

In contrast to deconvolution methods, the reverberation suppression method based on spectral subtraction [9] is not sensitive to the fluctuation of the impulse responses. The method estimates the power spectrum of the reverberation and then subtracts it from the power spectrum of the reverberant speech. The problem in the spectral subtraction is the nonlinear processing distortion caused by oversubtraction of the reverberation. The distortion degrades the quality of the processed reverberant speech.

This paper proposes a combination of MINT-based blind deconvolution and modified spectral subtraction for suppressing the tail of reverberation and improving the processed speech quality. MINT inverse filtering reduces the early reflection that has most of the power of the reverberation, and then, the modified spectral subtraction suppresses the tail of the inverse-filtered reverberation. Inverse filtering makes the power of the reverberation small, so the nonlinear processing distortion of spectral subtraction is reduced with a small subtraction of the power.

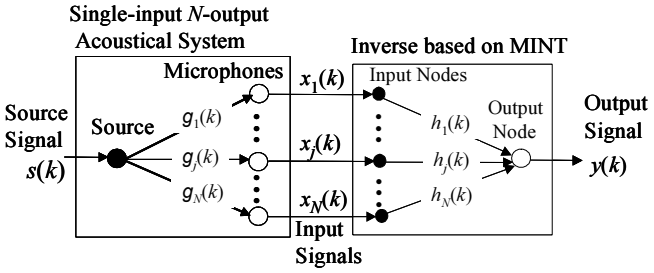

Fig. 1. MINT inverse filtering framework for single-input N-output acoustical system.

2. BLIND DECONVOLUTION BASED ON MINT INVERSE FILTERING

Consider a single-input N-output acoustical system shown in Fig. 1. Let $s(k)$ represent a source signal, and $x_j(k)$ represent the signal received at the jth microphone. Moreover, let $y(k)$ represent the inverse-filtered signal of $s(k)$. $g_j(k)$ denotes the impulse responses of the acoustic signal-transmission channel between the source and jth output of the system. $h_j(k)$ denotes the impulse response of an FIR filter connected to the jth output of the system.

The MINT inverse filtering of the system can be defined by the expression

$$\begin{aligned} \mathbf{B} &= \mathbf{GH}, \\ \mathbf{B} &= \begin{bmatrix} 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 & \cdots & \mathbf{G}_N \end{bmatrix}, \\ \mathbf{H} &= \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \vdots \\ \mathbf{h}_j \\ \vdots \\ \mathbf{h}_N \end{bmatrix}, \mathbf{h}_j = \begin{bmatrix} h_j(0) \\ h_j(1) \\ \vdots \\ h_j(L-1) \end{bmatrix}, \end{aligned} \tag{1}$$

$$\mathbf{G}_j = \begin{bmatrix} g_j(0) & 0 & \cdots & 0 \\ g_j(1) & g_j(0) & \cdots & \vdots \\ \vdots & g_j(0) & \ddots & 0 \\ g_j(K-1) & \vdots & \ddots & g_j(0) \\ 0 & g_j(K-1) & \ddots & g_j(1) \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & g_j(K-1) \end{bmatrix},$$

where $\mathbf{B}$: $NL \times 1$ target vector, $\mathbf{G}$: $NL \times NL$ impulse response matrix, $\mathbf{G}_j$ denotes the jth column of matrix $\mathbf{G}$, $\mathbf{H}$: $NL \times 1$ inverse filter vector, K: the length of the impulse response, and L: the length of the inverse filter. According to MINT [1], if there are no common zeros between the transfer functions of the impulse responses, the desired source signal can be recovered by inverse filtering.

The conventional MINT method uses room impulse responses to calculate the inverse, so it cannot recover speech signals in the practical situation where the room impulse responses are unknown in advance. However, the correlation matrix between received signals, which contains information about impulse responses, is available to the user. MINT-based inverse filters can be computed using this correlation matrix [8].

The correlation matrix of the received signals is defined by

$$\begin{aligned} \mathbf{R} &= E\{\mathbf{X}^{\mathrm{T}}\mathbf{X}\} \\ &= \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \cdots & \mathbf{R}_{1N} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \cdots & \mathbf{R}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{N1} & \mathbf{R}_{N2} & \cdots & \mathbf{R}_{NN} \end{bmatrix}, \end{aligned} \tag{2}$$

where $\mathbf{R}$: $NL \times NL$ correlation matrix, $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2\ \cdots \mathbf{X}_N]$, $\mathbf{X}_i = [x_i(k)\ x_i(k-1) \cdots x_i(k-(L-1))]$, $E\{\cdot\}$: expectation, and T: transpose.

We assume that the source signal is statistically white. That is

$$E\{s(k)s(k+n)\} = \begin{cases} \delta(n) & n = 0 \\ 0 & n \neq 0. \end{cases} \tag{3}$$

Using (3), the relationship between $\mathbf{R}$ and $\mathbf{G}$ is given by

$$\mathbf{R} = \mathbf{G}^{\mathrm{T}}\mathbf{G}. \tag{4}$$

Although the speech signal is not statistically white, it is modeled as a convolution of the white signal $s(k)$ and minimum phase filter $a(k)$ that has the characteristic of a long-term averaged speech spectrum. We use whitening filter $a^{-1}(k)$ to remove correlation due to speech, where $a(k) * a^{-1}(k) = \delta(k)$. $a(k)$ is estimated by averaging the power spectrum of the received signals.

Here, we also assume that the first microphone ($j = 1$) is closest to the source; i.e.,

$$g_j(0) = \begin{cases} g_1(0) & j = 1 \\ 0 & j \neq 1. \end{cases} \tag{5}$$

Multiplying $\mathbf{G}^{\mathrm{T}}$ by $\mathbf{B}$ yields

$$\mathbf{G}^{\mathrm{T}}\mathbf{B} = g_1(0)\mathbf{B}. \tag{6}$$

Finally, the MINT inverse filter $\mathbf{H}$ is obtained from (1), (4), and (6), and is given by

$$\mathbf{H} = g_1(0)\mathbf{R}^{-1}\mathbf{B}. \tag{7}$$

The term $g_1(0)$ in (7) is a scaling factor of the inverse. Although its value is unknown, we can set $g_1(0)$ to an arbitrary constant because scaling is not important in computing the inverse. The deconvolved signal $y(k)$ is given by inverse filtering the received signal $x_j(k)$.

3. MODIFIED SPECTRAL SUBTRACTION FOR SUPPRESSING LATE REVERBERATION

The deconvolution based on inverse filtering does not improve the tail part of reverberation because impulse responses are always fluctuating in the real world and the estimation error of inverse filters is caused by deviation of the correlation matrix averaged for a finite duration. The reverberation suppression method based on spectral subtraction was introduced by Lebart and Boucher [9]. The method estimates the power spectrum of the reverberation and then subtracts it from the power spectrum of reverberant speech. They modeled the impulse response as the outcome of the nonstationary random process with an exponential decay function to estimate the power of the reverberation. However, the deconvolved impulse responses do not exhibit the exponential decay, so we use a different model.

We modify the conventional spectral subtraction to combine with MINT inverse filtering for the suppression of the late reverberation. We assume that the short-time Fourier transform (STFT) $Y(\omega, m)$ of inverse-filtered speech $y(k)$ is a linear combination of the STFT $S(\omega, m)$ of original speech $s(k)$, that is

$$Y(\omega, m) = S(\omega, m) + \sum_{i=1}^{M} \alpha_i(\omega) S(\omega, m-i), \tag{8}$$

where indexes ω and m refer to frequency bin and time frame, respectively, $\alpha_i(\omega)$ is the coefficient of the late reverberation for previous i frames, and M is the duration of the reverberation.

Here, $\alpha_i(\omega) \ll 1$ because the inverse filtering reduces the early reflection part that has most of the power of the reverberation. Therefore, the power spectrum of late reverberation can be approximated by

$$\begin{aligned} P(\omega, m) &= \sum_{i=1}^{L-1} |\alpha_i(\omega)|^2 |S(\omega, m-i)|^2 \\ &\approx \sum_{i=1}^{L-1} |\alpha_i(\omega)|^2 |Y(\omega, m-i)|^2. \end{aligned} \tag{9}$$

Assuming the reverberation components are approximately uncorrelated between frames, the coefficients of the late reverberation are estimated by

$$\alpha_i(\omega) = E\left\{ \frac{|Y(\omega, m) Y^*(\omega, m-i)|}{|Y(\omega, m-i)|^2} \right\}. \tag{10}$$

Spectral subtraction is employed to estimate the original speech:

$$Z(\omega, m) = G(\omega, m) Y(\omega, m), \tag{11}$$

where $Z(\omega, m)$ is the STFT of recovered speech $z(k)$,

$$G(\omega, m) = \left\{ \frac{|Y(\omega, m-i)|^2 - P(\omega, m)}{|Y(\omega, m-i)|^2} \right\}^{1/2}, \tag{12}$$

and if $G \leq 0$ then $G = 0$ or a small constant value. The dereverberated signal $z(k)$ is reconstructed from the estimated STFT $Z(\omega, m)$, through the inverse-STFT and overlap-add techniques.

4. IMPLEMENTATION

We describe the overview of the complete algorithm of the proposed method in this section. The signal flow of the proposed method from the speech source to the recovered signal is shown in Fig. 2.

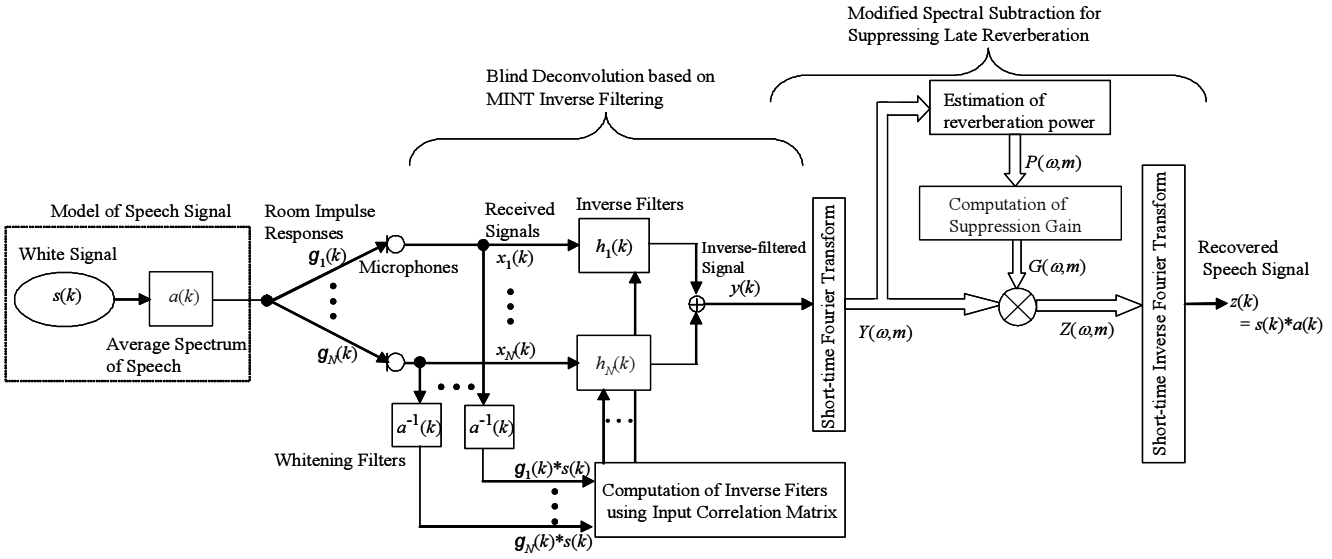

Fig. 2. Signal flow of proposed method.

Here, the speech signal is modeled as the convolution of white signal $s(k)$ and long-term averaged spectrum $a(k)$ and represented as $s(k) * a(k)$. The speech signal is reverberated by the room impulse responses $g_j(k)$ and received by the microphones. The received signals $x_j(k)$ are convolved by the whitening filter $a^{-1}(k)$ to remove the correlation due to speech and estimate the correlation matrix. The inverse filters $h_j(k)$ are computed by (7). The inverse-filtered signal $y(k)$ is obtained by convolving $x_j(k)$ with $h_j(k)$ and mixing these convolved signals. $y(k)$ is analyzed by the STFT into frequency components $Y(\omega, m)$. The power, $P(\omega, m)$, of the reverberation is estimated by (9). The suppression gain, $G(\omega, m)$, is calculated by (12). $Y(\omega, m)$ multiplied by $G(\omega, m)$ is the frequency components $Z(\omega, m)$ of the dereverberated signal $z(k)$. An inverse STFT is performed on $Z(\omega, m)$ to recover $z(k)$. This algorithm has been implemented on a Pentium IV 2.8 GHz Windows computer with audio interfaces for the real-time dereverberation.

5. EXPERIMENTS

The experimental results of objective and subjective evaluation are provided in the following to demonstrate the performance of the proposed method for speech dereverberation.

5.1. Dereverberation results of speech and impulse signals

In experiments, reverberated speech signals were obtained by convolution of anechoic phrases by real room impulse responses that were measured by an omnidirectional 4-microphone array spaced with a source-receiver distance of 3.8 m and the distance between microphones is 0.07 m. The dimensions of the room are 6.6 × 4.6 × 3.1 m, and the reverberation time is 0.55 s. The signals were sampled at 12 kHz, and the frame size is 1024 samples with a 256-sample-frame shift in the spectral subtraction. The length of inverse filter L is 2000 taps, the length of the whitening filter is 512 taps, and the duration for averaging the correlation matrix is 10 s,

For evaluating the effect of the inverse filtering, the inverse filters were estimated from the reverberant speech signal, and an impluse signal was deconvolved instead of the speech signal. As shown in Fig. 3(b), the reverberation in the inverse-filtered impulse response was suppressed well in comparison with the original room impulse response shown in Fig. 3(a). The waveform of the Japanese word '*son-na*', in the anechoic condition is shown in Fig. 4(a). Reverberant speech, inverse-filtered speech, and speech dereverberated by the proposed method are shown in panels in Figs. 4(a), (b), and (c), respectively. In Fig. 4(c), we can see that the reverberation was reduced and the pitch pulses of speech were recovered by inverse filtering. However, the tail part of the reverberation caused by the estimation error remained in the inverse-filtered signal. Comparing Fig. 4(d) to Fig. 4(c), the suppression of the reverberation tail is noteworthy in the signal recovered by the proposed method.

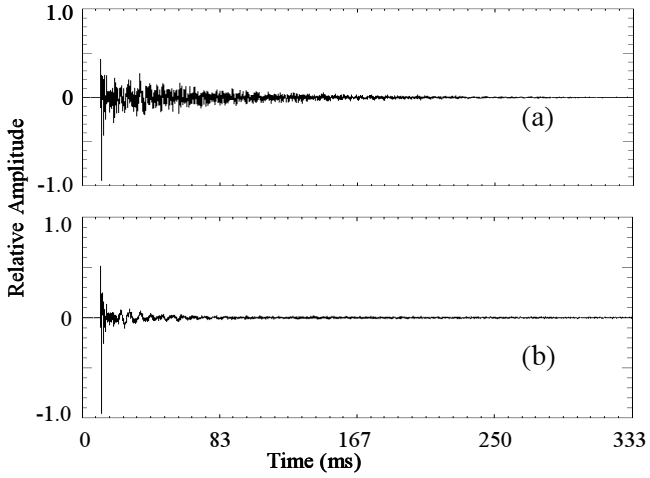

Fig. 3. Dereverberation result of impulse signal: (a) original room impulse response; (b) inverse-filtered impulse response.

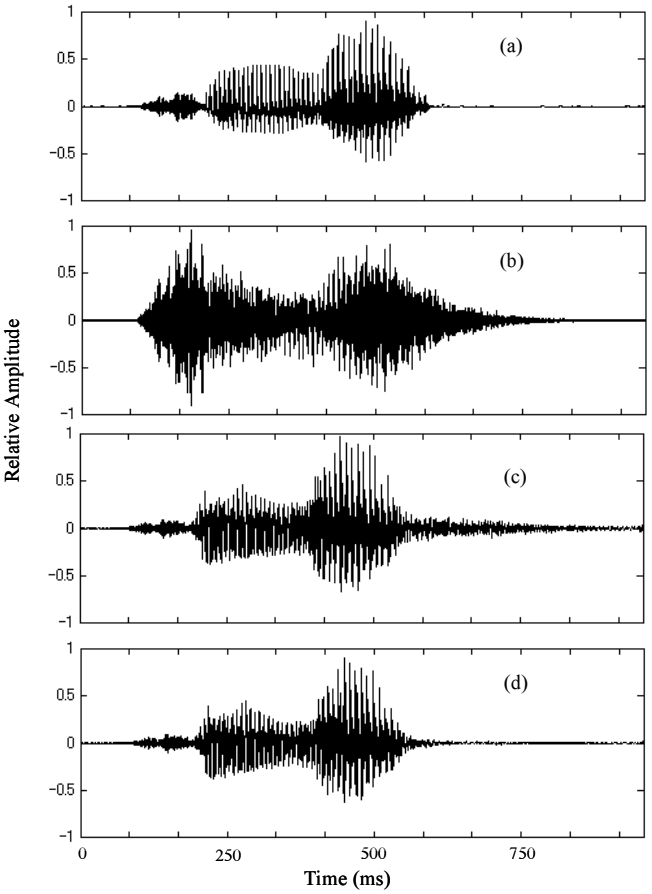

Fig. 4. Waveformes of speech signals: (a) original signal; (b) signal degraded by reverberation; (c) signal inverse-filtered by MINT-based blind deconvolution; and (d) signal recovered by proposed method combining inverse filtering and modified spectral subtraction.

5.2. Subjective assessment of the processed speech quality

We compared the proposed method with conventional spectral subtraction, from the viewpoint of subjective quality under the conditions where the speech signals are 3 male and 3 female voices. The assessment method is the comparison category rating (CCR) method [10]. The subjects are 24 non-experts. The inverse-filtered speech was included for comparison. Clean speech and reverberant speech were included as anchors.

The assessment results are shown in Table 1. In the CCR method, the second sample has better quality than the first if the CMOS (Comparison Mean Opinion Score) is more than 0. The results indicate that the proposed method provided better quality than conventional spectral subtraction. The score of the proposed method gave the best improvement in the quality in comparison with the reverberant speech except the original speech. There is no significant difference in quality between the inverse filtering and the conventional spectral subtraction.

6. CONCLUSION

We proposed a blind dereverberation method combining MINT blind deconvolution and modified spectral subtraction. MINT inverse filtering reduces early reflection, which has most of the power of the reverberation, and then spectral subtraction suppresses the tail of the inverse-filtered reverberation. The algorithm of the proposed method was implemented on a computer with audio interfaces for real-time speech derevberation. Dereverberation experiments demonstrated that the proposed method is effective in removing reverberation and improves the quality of reverberant speech and conventional spectral subtraction.

Table 1. Subjective CMOS (Comparison Mean Opinion Score). The CCR rating categories were used that the second sample compared to the first is 3: 'Much Better', 2: 'Better', 1: 'Slightly Better', 0: 'About the Same', -1: 'Slightly Worse', -2: 'Worse', -3: 'Much Worse'.

Condition (first sample vs. second sample)	CMOS	95% confidence interval
Subtraction vs. Proposed	0.44	0.13
Reverberant vs. Proposed	1.12	0.12
Reverberant vs. Inverse-filter	0.64	0.13
Reverberant vs. Subtraction	0.66	0.11
Reverberant vs. Original	1.97	0.11

7. REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, Vol. 36, No. 2, pp. 145-152, 1988.

[2] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, Vol. 66, No.1, pp. 165-169, 1979.

[3] M. I. Gurelli and C. L. Nikias, "EVAM: an eigenvector-based algorithm for multi-channel blind deconvolution of input colored signals," *IEEE Trans. SP*, Vol. 43, No. 1, pp. 134-149, 1995.

[4] H. Wang, "Multi-channel deconvolution using Pade approximation," *Proc. of the ICASSP 95*, pp. 3007-3010, Detroit, U.S.A., Apr. 1995.

[5] A. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, No. 7, pp. 1129-1159, 1995.

[6] K. Furuya and Y. Kaneda, "Two-channel blind deconvolution for non-minimum phase impulse responses," *Proc. of the ICASSP 97*, pp. 1315-1318, Munchen, Germany, Apr. 1997.

[7] T. Hikichi, M. Delcroix, and M. Miyoshi, "Blind dereverberation based on estimates of signal transmission channels without precise information on channel," *Proc. of the ICASSP 2005*, pp. 1069-1072, Mar. 2005.

[8] K. Furuya, "Noise reduction and dereverberation using correlation matrix based on the multiple-input/output inverse-filtering theorem (MINT)," *Proc. of International Workshop on Hands-free Speech Communication*, pp. 59-62, Japan, Apr. 2001.

[9] K. Lebart and J. M. Boucher, "A new method based on spectral subtraction for speech dereverberation.," *Acta Acoustica*, Vol. 87, pp. 359-366, 2001.

[10] ITU-T Recomendation 800 Annex E, "Comparison Category Rating (CCR) method," 1996.