HYBRID DEREVERBERATION USING BLIND DECONVOLUTION AND SPECTRAL SUBTRACTION TO COMPENSATE FOR MOTION OF SOURCE

Ken'ichi Furuya and Akitoshi Kataoka

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

ABSTRACT

A hybrid dereverberation method for speech enhancement in a situation requiring adaptation where a speaker shifts his head and impulse responses are frequently changed under reverberant conditions is presented. We combine MINT-based blind deconvolution with modified spectral subtraction of the estimation error of inverse filters obtained in practice. Our method computes inverse filters by estimating the correlation matrix between input signals that can be observed, without measuring room impulse responses. The transient performance of the proposed method in the adaptation is demonstrated with experiments using measured room impulse responses.

1. INTRODUCTION

When a speaker is some distance away from a microphone in a teleconference, the speech signal is distorted by room reverberation, so it is less intelligible to listeners. One theoretical method to achieve almost complete 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 approach to measure the impulse responses between the speaker and the microphones.

A number of multichannel blind deconvolution methods, [3] – [8], that do not measure room impulse responses have recently been developed for speech dereverberation. 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 are not effective at reducing the tail of reverberation in the actual world.

In contrast to deconvolution methods, the reverberation suppression method based on spectral subtraction [10] is not sensitive to the fluctuation of 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 spectral subtraction is the nonlinear processing distortion, for example in the case of musical noise, caused by over-subtraction of the reverberation. The distortion degrades the quality of the processed reverberant speech.

We propose a hybrid dereverberation method by combining MINT-based blind deconvolution and modified spectral subtraction for suppressing the tail of reverberation and improving the processed speech quality [9]. MINT inverse filtering reduces the early reflection that constitutes most of the power of the reverberation. Then, the modified spectral subtraction suppresses the tail of the inverse-filtered reverberation. Inverse filtering reduces the power of the reverberation, so the nonlinear processing distortion of spectral subtraction is reduced with a small subtraction of the power. In this work, we go a step further and expand the hybrid dereverberation to a situation requiring adaptation where a speaker shifts his head and impulse responses are frequently changed. Inverse filters are adjusted by using the exponentially time-averaged correlation matrix in which recent components are emphasized and older components fade out. The transient performance of the proposed adaptive method is investigated in objective and subjective experiments.

2. HYBRID DEREVERBERATION

To obtain better dereverberation properties, we use a hybrid processing scheme that works in two sequential stages, performing the following.

- MINT-based blind deconvolution: reverberant speech signal is blindly inverse-filtered by using an exponentially time-averaged correlation matrix.
- Modified spectral subtraction: spectral subtraction is applied to suppress late reverberation.

2.1. Blind Deconvolution Based on MINT Inverse Filtering

2.1.1. Review of Conventional MINT Inverse Filtering

The inverse of the single-input single-output acoustical system becomes unstable because the acoustic signal-transmission channel is generally considered to be nonminimum phase. Miyoshi and Kaneda proposed the MINT method for achieving an exact inverse of an acoustic system [1]. Using MINT, the inverse is constructed from multiple FIR (Finite Impulse Response) filters by adding acoustic signal-transmission channels produced by using multiple microphones.

Consider a single-input N-output acoustical system. 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 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.

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

$$\mathbf{B} = \mathbf{G}\mathbf{H},$$

$$\mathbf{B} = \begin{bmatrix} 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_2 & \cdots & \mathbf{G}_N \end{bmatrix}, \tag{1}$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \vdots \\ \mathbf{h}_j \\ \vdots \\ \mathbf{h}_N \end{bmatrix}, \ \mathbf{h}_j = \begin{bmatrix} h_j(0) \\ h_j(1) \\ \vdots \\ h_j(L-1) \end{bmatrix},$$

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

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

$$\mathbf{H} = \mathbf{G}^{-1}\mathbf{B}. \tag{2}$$

2.1.2. Computation of Inverse Filter Using Correlation Matrix

The conventional MINT method uses room impulse responses to calculate the inverse filter, so it cannot recover speech signals in a 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 received signals is defined by

$$\begin{aligned} \mathbf{R} &= E\{\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k\} \\ &= \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \cdots & \mathbf{R}_{1N} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \cdots & \mathbf{R}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{N1} & \mathbf{R}_{N2} & \cdots & \mathbf{R}_{NN} \end{bmatrix}, \end{aligned} \tag{3}$$

where $\mathbf{R}$ is the $NL \times NL$ correlation matrix,
$\mathbf{X}_k = [\mathbf{X}_{1k}, \ \mathbf{X}_{2k}, \ \cdots, \mathbf{X}_{Nk}]$,
$\mathbf{X}_{ik} = [x_i(k) \ x_i(k-1), \cdots, x_i(k-(L-1))]$,
$E\{\cdot\}$ is the expectation, and T is the transpose.

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

$$E\{s(k)s(k+n)\} = \delta(n) \tag{4}$$

Using (4), the relationship between $\mathbf{R}$ and $\mathbf{G}$ is given by

$$\mathbf{R} = \mathbf{G}^{\mathrm{T}}\mathbf{G}. \tag{5}$$

Although the speech signal is not statistically white, it is modeled as a convolution of the white signal $s(k)$ and the minimum phase filter $a(k)$. $a(k)$ 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 received signals.

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

$$g_j(0) = \begin{cases} g_1(0) & j = 1 \\ 0 & j \neq 1. \end{cases} \tag{6}$$

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

$$\mathbf{G}^{\mathrm{T}}\mathbf{B} = g_1(0)\mathbf{B}. \tag{7}$$

Finally, MINT inverse filter $\mathbf{H}$ is obtained from (2), (5), and (7), and is given by

$$\mathbf{H} = g_1(0)\mathbf{R}^{-1}\mathbf{B}. \tag{8}$$

The term $g_1(0)$ in (8) 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, $x_j(k)$.

2.1.3. Computation of Inverse Filter in Situation Requiring Adaptation

In the situation requiring adaptation where impulse response $\mathbf{G}$ is frequently changed, we use the following recursive time-averaging to estimate the correlation matrix instead of using (3):

$$\hat{\mathbf{R}}_k = \beta\hat{\mathbf{R}}_{k-1} + (1-\beta)\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k, \tag{9}$$

where $\hat{\mathbf{R}}_k$ is the estimate of $\mathbf{R}$ at k, β is the weight of the older estimate $\hat{\mathbf{R}}_{k-1}$ in the time-averaging. As a rule of thumb, we could let β be chosen such that the half-life of the exponential function is equal to the value of the duration over which $\mathbf{G}$ is stationary. Using (9), inverse filter $\hat{\mathbf{H}}_k$ at k in the adaptive situation is obtained and given by

$$\hat{\mathbf{H}}_k = g_1(0)\hat{\mathbf{R}}_k^{-1}\mathbf{B}. \tag{10}$$

2.2. Modified Spectral Subtraction for Suppressing Late Reverberation

The deconvolution based on inverse filtering does not improve the tail 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 [10]. 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 an outcome of the nonstationary random process using an exponential decay function to estimate the power of the reverberation. However,

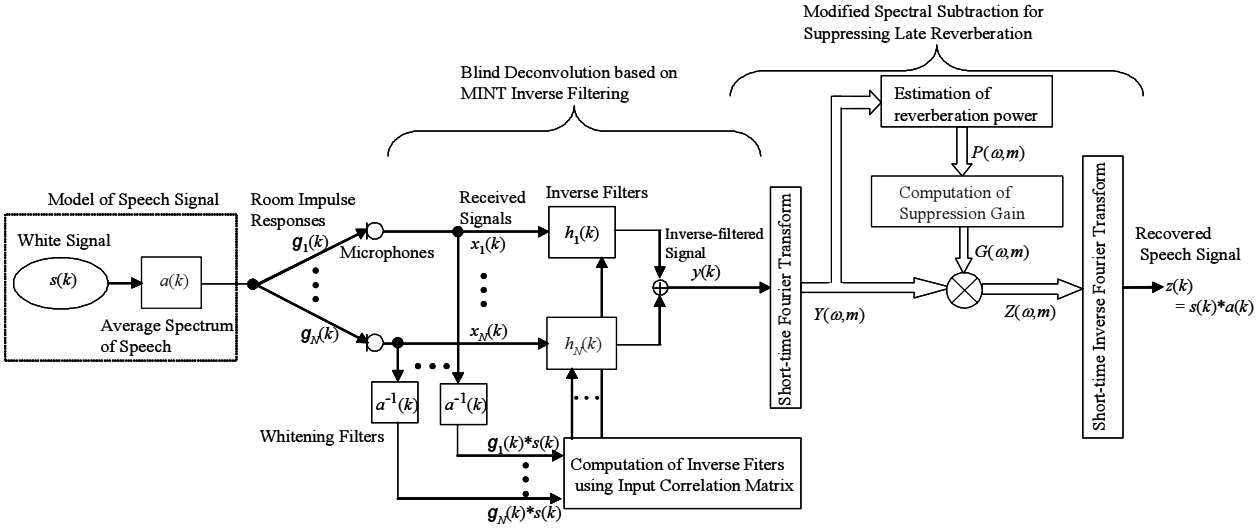

Figure 1: *Signal flow of proposed method.*

the deconvolved impulse responses do not exhibit exponential decay, so we use a different model.

We modify conventional spectral subtraction to combine it with MINT inverse filtering for the suppression of 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 $s(k)$, which is

$$Y(\omega, m) = S(\omega, m) + \sum_{i=1}^{M} \alpha_i(\omega) S(\omega, m-i), \qquad (11)$$

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 inverse filtering reduces the early reflection part that constitutes most of the power of the reverberation. Therefore, the power spectrum of late reverberation can be approximated by

$$\begin{aligned} P(\omega, m) &= \sum_{i=1}^{M} |\alpha_i(\omega)|^2 |S(\omega, m-i)|^2 \\ &\approx \sum_{i=1}^{M} |\alpha_i(\omega)|^2 |Y(\omega, m-i)|^2. \end{aligned} \qquad (12)$$

Assuming the reverberation components $\alpha_i(\omega) S(\omega, m-i)$ in (11) are weakly correlated between frames i and $\alpha_i(\omega) \ll 1$, the coefficients of the late reverberation are estimated by

$$\begin{aligned} \alpha_i(\omega) &= E\left\{ \frac{Y(\omega, m) S^*(\omega, m-i)}{|S(\omega, m-i)|^2} \right\} \\ &\approx E\left\{ \frac{Y(\omega, m) Y^*(\omega, m-i)}{|Y(\omega, m-i)|^2} \right\}. \end{aligned} \qquad (13)$$

Spectral subtraction is used to estimate the original speech:

$$Z(\omega, m) = G(\omega, m) Y(\omega, m), \qquad (14)$$

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

$$G(\omega, m) = \left\{ \frac{|Y(\omega, m)|^2 - P(\omega, m)}{|Y(\omega, m)|^2} \right\}, \qquad (15)$$

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.

3. IMPLEMENTATION

We describe an 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. 1. 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 room impulse responses $g_j(k)$ and received by microphones. Received signals $x_j(k)$ are convolved by whitening filter $a^{-1}(k)$ to remove the correlation due to speech and estimate the correlation matrix. Inverse filters $h_j(k)$ are computed using (10). 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 using (12). The suppression gain, $G(\omega, m)$, is calculated using (15). $Y(\omega, m)$ multiplied by $G(\omega, m)$ gives 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.

4. EXPERIMENTS

Experimental results of objective and subjective evaluations are provided below to demonstrate the transient performance of the proposed method for speech dereverberation.

4.1. Transient performance of inverse filtering in adaptation

In experiments, reverberated speech signals were obtained by convolution of anechoic phrases and real room impulse responses that were measured by an omnidirectional 4-microphone array

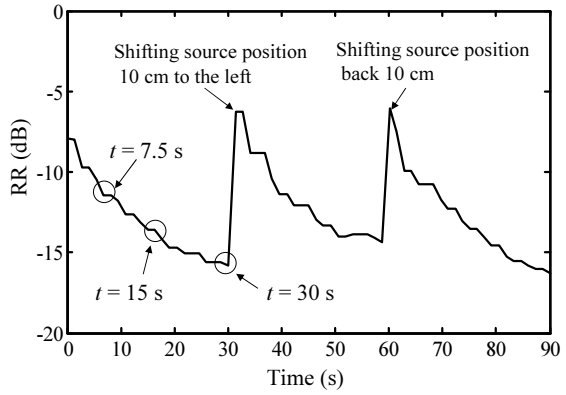

Figure 2: *Power of reverberation in inverse-filtered impulse response (RR) in adaptive situation where source position was shifted 10 cm to the left at t =30 s and shifted back 10 cm at t = 60 s.*

spaced at a source-receiver distance of 3.8 m. The distance between microphones is 7 cm. The source position was shifted 10 cm to the left at $t = 30$ s and shifted to back 10 cm at $t = 60$ s. The dimensions of the room are $6.6 \times 4.6 \times 3.1$ m, and the reverberation time is 0.55 s. The signals were sampled at 12 kHz, and the frame size is 1,024 samples with a 512-sample-frame shift in the spectral subtraction. The length of inverse filter L is 2,048 taps, the length of the whitening filter is 512 taps, and β is 0.5 s. For evaluating the effect of the inverse filtering, the inverse filters were estimated from the reverberant speech signal, and an impulse signal was deconvolved instead of the speech signal. The power of reverberation in the inverse-filtered impulse response (RR) was used as an inverse-filtering performance measure. RR is defined as

$$\mathrm{RR} = 10 \log_{10} \frac{\int_{\tau_0}^{\infty} \gamma^2(\tau) d\tau}{\int_{0}^{\infty} \gamma^2(\tau) d\tau}, \qquad (16)$$

where $\gamma(\tau)$ is the inverse-filtered impulse response and τ_0 is the time boundary between direct sound and reverberation of the impulse response, where $\tau_0 = 50$ ms. As shown in Fig. 2, inverse filtering adaptively reduced the power of reverberation and tracked changes of the source position.

4.2. Subjective assessment of processed speech quality

We compared the proposed method with conventional inverse filtering and spectral subtraction from the viewpoint of subjective quality. Inverse-filtered speech signals at $t = 7.5$, 15, and 30 s in Fig. 2 were included for evaluating transient speech quality. Signals dereverberated by the proposed method were obtained by performing the modified spectral subtraction on the inverse-filtered speech signals. The speech signals were of 3 male and 3 female voices. The assessment method was the absolute category rating (ACR) method [11]. Subjects were twenty four non-experts. Clean speech and reverberant speech were included as anchors. The assessment results are shown in Table 1. These results indicate that the proposed method provided better speech quality than conventional inverse filtering and spectral subtraction at every point. The score of the proposed method gave the best improvement in quality in comparison with reverberant speech except the original speech.

Table 1: *Subjective MOS (Mean Opinion Score). ACR rating categories were 5: 'Excellent', 4: 'Good', 3: 'Fair', 2: 'Poor', and 1: 'Bad'.*

Condition	MOS	95% confidence interval
Clean speech	4.40	0.12
Reverberant speech	2.30	0.14
Spectral subtraction	2.78	0.15
Inverse-filtering at $t = 7.5$ s	2.33	0.11
at $t = 15$ s	2.50	0.14
at $t = 30$ s	2.75	0.16
Proposed method at $t = 7.5$ s	2.92	0.15
at $t = 15$ s	3.40	0.14
at $t = 30$ s	3.70	0.14

5. CONCLUSION

We proposed a hybrid dereverberation method for speech enhancement in the adaptive situation where a speaker shifts his head and impulse responses are frequently changed under reverberant conditions. MINT-based blind deconvolution was combined with modified spectral subtraction of the estimation error of inverse filters in the field to improve inverse-filtered speech quality. The algorithm of the proposed method was implemented on a computer with audio interfaces for real-time speech dereverberation. Dereverberation experiments demonstrated that the proposed method is effective in the situation requiring adaptation and improves the quality of reverberant speech, conventional inverse filtering and spectral subtraction.

6. 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, 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. Furuya, S. Sakauchi, and A. Kataoka, "Speech dereverberation by combining MINT-based blind deconvolution and modified spectral subtraction," *Proc. of the ICASSP 2006*, Vol. 1, pp. 813-816, May 2006.

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

[11] ITU-T Recomendation 800 Annex B, "Absolute Category Rating (ACR) method," 1996.