Noise robust speech dereverberation using constrained inverse filter

Ken’ichi Furuya\textsuperscript{1}, Akitoshi Kataoka\textsuperscript{2}, and Yoichi Haneda\textsuperscript{1}

\textsuperscript{1}NTT Cyber Space Laboratories, Nippon Telegraph and Telephone Corporation, Tokyo, Japan
\textsuperscript{2}Faculty of Science and Technology, Ryukoku University, Shiga, Japan

Abstract

A noise robust dereverberation method is presented for speech enhancement in noisy reverberant conditions. This method introduces the constraint of minimizing the noise power in the inverse filter computation of dereverberation. It is shown that there exists a tradeoff between reducing the reverberation and reducing the noise: this tradeoff can be controlled by the constraint. Inverse filtering reduces early reflections and directional noise. In addition, spectral subtraction is used to suppress the tail of the inverse-filtered reverberation and residual noise. The performance of our method is objectively and subjectively evaluated in experiments using measured room impulse responses. The results indicate that this method provides better speech quality than the conventional methods.

Index Terms: dereverberation, noise reduction, blind deconvolution, spectral subtraction, speech enhancement

1. Introduction

When a speaker is some distance away from the microphone in a hands-free teleconference, the speech signal is distorted by room reverberation, so it is less intelligible to listeners. Thus, dereverberation is an important speech enhancement process for hands-free communication. A number of speech dereverberation methods \cite{1} – \cite{9} have recently been developed. We introduced a hybrid dereverberation method that combines multichannel blind deconvolution and spectral subtraction \cite{10}. This hybrid method provides superior speech quality to conventional methods for noise-free conditions. However, conference rooms usually have reverberation together with noise sources, such as computer fans or air conditioning systems, and these also degrade the speech quality. It is therefore necessary to develop dereverberation algorithms that also perform noise reduction.

Several dereverberation methods for the noisy reverberant condition have been proposed. They can be considered in two categories: (i) multichannel blind deconvolution for a multi-source scenario, i.e., a target source and one or more noise sources \cite{2}, \cite{3}, \cite{4}, and (ii) Wiener postfiltering or spectral subtraction for statistical models of reverberation and noise \cite{8},\cite{9}. Blind deconvolution methods are generally not so robust against incoherent noise, which is not spatially localized, and are not effective at reducing the late reverberation and diffuse noise in the actual world. The problem in spectral subtraction is the nonlinear processing distortion, caused by over-subtraction of the reverberation or noise. This distortion degrades the total quality of the processed reverberant speech.

In this paper, we propose a noise robust dereverberation method for speech enhancement in noisy reverberant conditions. We introduce the constraint of minimizing the noise power in the inverse filter computation of blind deconvolution.

We show that there exists a tradeoff between reducing the reverberation and reducing the noise. This tradeoff can be controlled using the constraint. The blind deconvolution with the constrained inverse filter reduces the early reflection component and directional noise component. In addition, spectral subtraction is used to suppress the late reverberation component and residual noise component. The blind deconvolution reduces the power of the reverberation and noise, so the nonlinear processing distortion of spectral subtraction is reduced using a small subtractive power.

2. Speech dereverberation using noise-constrained inverse filter

In this section, we give an overview of the speech dereverberation using noise-constrained inverse filter. Suppose that we have a speech source, noise sources, and \(N\) microphones in a room, as shown in Fig. 1. Here, the speech signal \(s(k)\) is reverberated by room impulse responses \(g_j(k)\) and received with the noise signal \(q_j(k)\) by the \(j\)th microphone as

\[
x_j(k) = s(k) * g_j(k) + q_j(k),
\]

where \(k\) refers to the time sample, and \(j\) is the index of the microphone \((j = 1, 2, \ldots, N)\). The noise signals \(q_j(k)\) are assumed to be statistically stationary and to be characterized by coherent components (direct paths and first reflections of the localized noise sources) and by incoherent or diffuse noise components (modifications of the noise from noise sources by the acoustical environment). The inverse-filtered signal \(y(k)\) is obtained by convolving the received signals \(x_j(k)\) with the inverse filters \(h_j(k)\) and mixing these convolved signals.

\[
y(k) = \sum_{j=1}^{N} x_j(k) * h_j(k)
\]

2.1. Inverse filter of dereverberation in noise-free condition

First, the inverse filter of dereverberation in the noise-free condition, that is \(q_j(k) = 0\), is derived. The MINT \cite{12} inverse...
filtering of the system can be defined by the expression

\[ \mathbf{b} = \mathbf{Gh}, \]  

(3)

where \( \mathbf{b} \) is the output, \( \mathbf{G} \) is the impulse response matrix, and \( \mathbf{T} \) is the transpose.

\[ \mathbf{G}_j = \begin{bmatrix} g_j(0) & 0 & \cdots & 0 \\ g_j(1) & g_j(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_j(NL-1) & g_j(NL-2) & \cdots & g_j(0) \\ 0 & g_j(NL-1) & \cdots & g_j(NL-L) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_j(NL-L+1) \\ 0 & 0 & \cdots & g_j(NL-L) \\ \end{bmatrix}, \]

\[ \mathbf{h} = [h_1, h_2, \ldots, h_{NL}]^T, \]

\[ \hat{\mathbf{h}} = [h_1(0), h_2(1), h_{NL}(L-1)]^T, \]

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 \( j \)th \( NL \times L \) submatrix of \( \mathbf{G} \) and consists of \( g_j(k) \), the pure-delay common among the impulse response \( g_j(k) \) is eliminated, \( \mathbf{h} \) is the \( NL \times L \) inverse filter vector, \( K \) is the length of the impulse response, and \( L \) is the length of the inverse filter.

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 [2].

The correlation matrix of received signals is defined by

\[ \mathbf{R} = E\{\mathbf{X}_c^H \mathbf{X}_b\}, \]  

(4)

where \( \mathbf{R} \) is the \( NL \times NL \) correlation matrix, \( \mathbf{X}_c = [x_1, x_2, \ldots, x_{NL}] \), \( x_{ik} = [x_1(k), x_2(k-1), \ldots, x_{NL}(k-(L-1))] \), \( E\{\} \) is the expectation, and \( T \) is the transpose.

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

\[ E\{s(k)s(k+n)\} = \delta(n), \]  

(5)

where \( \delta(n) \) is the delta function. Using (5), we get the relationship between \( \mathbf{R} \) and \( \mathbf{G} \)

\[ \mathbf{R} = \mathbf{G}^H \mathbf{G}. \]  

(6)

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 \( a(k) \). This \( a(k) \) has the characteristics of a long-term averaged power spectrum. We use whitening filter \( a^{-1}(k) \) to remove correlation due to speech, where \( a(k) \) is a \( a^{-1}(k) = \delta(k) \). The \( a(k) \) is estimated by averaging the power spectrum of received signals, as described in [10].

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

\[ g_1(0) = \begin{cases} 0 & j = 1 \\ \varepsilon & j \neq 1. \end{cases} \]  

(7)

Then, multiplying \( \mathbf{G}^H \) by \( \mathbf{b} \) yields

\[ \mathbf{G}^H \mathbf{b} = g_1(0) \mathbf{b}. \]  

(8)

Finally, the estimate of the MINT inverse filter \( \hat{\mathbf{h}} \) is obtained from (3), (6), and (8). It is given by

\[ \hat{\mathbf{h}} = g_1(0) \mathbf{b}. \]  

(9)

If \( \mathbf{R} \) is nonsingular,

\[ \hat{\mathbf{h}} = g_1(0) \mathbf{R}^{-1} \mathbf{b}. \]  

(10)

The term \( g_1(0) \) in (10) 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. For convenience of computation, \( g_1(0) = 1 \) is used in practice.

2.2. Constraint for minimizing inverse-filtered noise power

The blind deconvolution using the inverse filter for the noise-free condition cannot reduce the noise component in the noisy environment shown in Fig. 1. Therefore, we introduce a soft constraint on computation, (10), which minimizes the noise power. Although the constrained inverse filter is different from the MINT inverse filter, it is desirable to attain maximum noise reduction while allowing a small degree of degradation in the dereverberation.

The expected power of the inverse filtered noise signal \( d(k) \) is given by

\[ \mathbb{E}\{d^2(k)\} = \mathbb{E}\{h^T \mathbf{Q}^T \mathbf{Q} \mathbf{h}\} = \mathbf{R}_Q \mathbf{h}, \]  

(11)

where \( \mathbf{R}_Q \) is the \( NL \times NL \) noise correlation matrix, \( \mathbf{Q}_k = E\{\mathbf{Q}_k \mathbf{Q}_k^H\} \), and \( \mathbf{q}_{nk} = [q_{1k}, q_{2k}, \ldots, q_{NLk}] \), and if the noise power is minimized, the gradient of (11) is zero:

\[ \frac{\partial \mathbb{E}\{d^2(k)\}}{\partial \mathbf{h}} = 2 \mathbf{R}_Q \mathbf{h} = 0. \]  

(12)

To obtain the constrained inverse filter that minimizes the noise power, we use the following simultaneous equations of (9) and (12) with weight \( \lambda \)

\[ \begin{bmatrix} \mathbf{R} \hfill \lambda \mathbf{R}_Q \hfill \end{bmatrix} \begin{bmatrix} \mathbf{h} \hfill \end{bmatrix} = \begin{bmatrix} \mathbf{b} \hfill \mathbf{0} \hfill \end{bmatrix}. \]  

(13)

Using the generalized inverse matrix, we can express the constrained inverse filter as

\[ \mathbf{h}(\lambda) = \left( \begin{bmatrix} \mathbf{R} \hfill \lambda \mathbf{R}_Q \hfill \end{bmatrix}^T \begin{bmatrix} \mathbf{R} \hfill \lambda \mathbf{R}_Q \hfill \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{R} \hfill \lambda \mathbf{R}_Q \hfill \end{bmatrix}^T \begin{bmatrix} \mathbf{b} \hfill \mathbf{0} \hfill \end{bmatrix}. \]  

(14)

We use the bursty nature of the speech signal and estimate the reverberant speech correlation matrix \( \mathbf{R} \) using the voice activity detection (VAD) when speech is present, and estimate the noise correlation matrix \( \mathbf{R}_Q \) when speech is absent. The noise component included in the estimate of \( \mathbf{R} \) can be cancelled by subtracting the noise correlation matrix.

The squared error equation of (13) can be expressed by

\[ \mathbf{e}^2(\lambda) = \begin{bmatrix} \mathbf{b} \hfill \mathbf{0} \hfill \end{bmatrix} - \begin{bmatrix} \mathbf{R} \hfill \lambda \mathbf{R}_Q \hfill \end{bmatrix} \mathbf{h}(\lambda)^T \mathbf{e}^2(\lambda) = \begin{bmatrix} \mathbf{b} \hfill \mathbf{0} \hfill \end{bmatrix} - \begin{bmatrix} \mathbf{R} \hfill \lambda \mathbf{R}_Q \hfill \end{bmatrix} \mathbf{h}(\lambda). \]  

(15)

where \( \mathbf{e}_C(\lambda) = \mathbf{b} - \mathbf{Rh}(\lambda), \mathbf{e}_Q(\lambda) = \mathbf{0} - \mathbf{R}_Q \mathbf{h}(\lambda). \) The quantities \( \mathbf{e}_C^2(\lambda) \) and \( \mathbf{e}_Q^2(\lambda) \) are the squared equation errors.
2.3. Spectral subtraction for suppressing late reverberation and residual noise

The deconvolution based on inverse filtering does not improve the late 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. In addition, the diffuse noise remains in speech after the deconvolution process because the multichannel deconvolution is a kind of beamforming to reduce the directional noise component which is coherent between the input channels and can hardly reduce the incoherent noise component. One way to suppress the late reverberation and residual noise is to perform spectral subtraction estimating original speech spectrum [7], [8], [10], [11]. Finally, we combine the deconvolution and spectral subtraction to obtain better speech quality. The deconvolution reduces the power of the reverberation and noise, so the nonlinear processing distortion of spectral subtraction is reduced using a small subtractive power.

3. Experiments

Here, we present experimental results that demonstrate the performance of the proposed method for speech enhancement in noisy reverberant conditions.

3.1. Experimental conditions

In the experiments, reverberant speech signals were obtained by convolving anechoic phrases and actual room impulse responses. The impulse responses were measured with an omnidirectional four-microphone array spaced such that the source-receiver distance was 3.8 m and the distance between microphones was 0.07 m, as shown in Fig. 2. The speech signals were of three male and three female voices. The length of each speech signal was 60 s. The dimensions of the room are 6.6 m × 4.6 m × 3.1 m and its reverberation time is 0.41 s. Hoth noise signal was used to simulate the noisy reverberant condition. The input power ratio of reverberant speech to noise was 10 dB at the first microphone. The signals were sampled at 12 kHz. The length of the inverse filter was 4096 taps. The frame size was 1024 samples with a 512-sample-frame shift in the spectral subtraction.

![Figure 2: Experimental conditions.](image)

![Figure 3: Spectrograms of speech signals in noisy reverberant conditions. (a) original clean speech, (b) noisy reverberant speech, (c) blind deconvolution without noise constraint, (d) conventional hybrid method, and (e) proposed method.](image)

3.2. Objective and subjective evaluation

The typical spectrograms of the processed speech signals are shown in Fig. 3. In Fig. 3(a), we see that the clean speech had a fine harmonic structure and silence gaps between the words. In contrast to the clean speech, the noisy reverberant speech was smeared and its harmonic structure was elongated, as shown in Fig. 3(b). Although blind deconvolution without noise constraint recovered the harmonic structure, it did not reduce noise, as shown in Fig. 3(c). The conventional hybrid method in [10] enables us to see the harmonic structure but the smearing remained, as shown in Fig. 3(d). Our method lessened the smearing and clarified silence gaps more than the other methods did, as shown in Fig. 3(e).

Objective evaluation was performed using the signal-to-interference ratio (SIR) and the noise-to-masking ratio (NMR) measures [10], [13]. When reverberation is considered as interference, the SIR is similar to the direct-to-reverberation ratio (DRR). The NMR is an objective measure based on the human auditory system and it indicates the audible noise components relative to the hearing threshold. Therefore, an NMR of 0 dB indicates a noise at the threshold of audibility, whereas higher NMRs mean more noticeable noise. This measure has been found to have a high degree of correlation with subjective tests [14]. In the computation of SIR and NMR, the desired signal is direct sound and the interference consists of both reverberation and noise. In addition, NMR_{reverb} and SIR_{reverb} were computed by picking up the reverberation component as the interference and used as measures to evaluate the dereverberation performance. NMR_{noise} and SIR_{noise} were computed by picking up only the noise component as the interference and used as measures to evaluate the noise reduction performance.

The NMR and SIR results are shown in Table 1, comparing noisy reverberant speech, spectral subtraction, blind...
Table 1: Results of objective and subjective evaluation experiments.

<table>
<thead>
<tr>
<th>Condition</th>
<th>NMR (dB)</th>
<th>NMR_\text{reverb}</th>
<th>NMR_\text{noise}</th>
<th>SIR (dB)</th>
<th>SIR_\text{reverb}</th>
<th>SIR_\text{noise}</th>
<th>MOS</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean speech</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Noisy reverberant speech</td>
<td>21.4</td>
<td>-1.52</td>
<td>21.0</td>
<td>5.12</td>
<td>9.71</td>
<td>16.0</td>
<td>12.4</td>
<td>9.88 21.0</td>
</tr>
<tr>
<td>Spectral subtraction</td>
<td>13.8</td>
<td>7.20</td>
<td>8.39</td>
<td>8.11</td>
<td>9.81</td>
<td>16.3</td>
<td>1.97</td>
<td>1.0 1.5</td>
</tr>
<tr>
<td>Blind deconvolution</td>
<td>21.1</td>
<td>1.94</td>
<td>20.5</td>
<td>5.60</td>
<td>16.0</td>
<td>6.12</td>
<td>2.28</td>
<td>1.0 1.1</td>
</tr>
<tr>
<td>without noise constraint</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conventional hybrid method</td>
<td>13.7</td>
<td>-1.35</td>
<td>10.1</td>
<td>9.88</td>
<td>11.3</td>
<td>13.6</td>
<td>2.37</td>
<td>1.0 1.1</td>
</tr>
<tr>
<td>Proposed method (\lambda = 1)</td>
<td>11.6</td>
<td>-1.50</td>
<td>7.10</td>
<td>11.2</td>
<td>12.1</td>
<td>18.3</td>
<td>2.44</td>
<td>1.3</td>
</tr>
<tr>
<td>(\lambda = 2)</td>
<td>9.10</td>
<td>-1.70</td>
<td>4.93</td>
<td>11.8</td>
<td>12.4</td>
<td>20.4</td>
<td>2.66</td>
<td>1.4</td>
</tr>
<tr>
<td>(\lambda = 5)</td>
<td>5.58</td>
<td>-1.22</td>
<td>0.66</td>
<td>11.9</td>
<td>12.2</td>
<td>23.8</td>
<td>2.78</td>
<td>1.4</td>
</tr>
<tr>
<td>(\lambda = 10)</td>
<td>5.27</td>
<td>-0.61</td>
<td>-0.08</td>
<td>11.6</td>
<td>11.9</td>
<td>23.5</td>
<td>2.82</td>
<td>1.5</td>
</tr>
</tbody>
</table>

We proposed a noise robust dereverberation method for speech enhancement in noisy reverberant conditions. It introduced the constraint of minimizing the noise power in the inverse filter computation. We showed that there exists a tradeoff between reducing the reverberation and reducing the noise and that this tradeoff can be controlled by the constraint. In addition, spectral subtraction was used to suppress the late reverberation component and residual noise component for simultaneous reduction of reverberation and noise. The performance of our method in noisy reverberant conditions was evaluated in experiments using measured room impulse responses. The results indicated that this method provided better speech quality than the conventional methods.

5. References