Blind Dereverberation Based on CMN and Spectral Subtraction by Multi-channel LMS Algorithm

L. Wang\(^1\), S. Nakagawa\(^2\), N. Kitaoka\(^3\)

\(^1\)Department of Systems Engineering, Shizuoka University, Japan
\(^2\)Department of Information and Computer Sciences, Toyohashi University of Technology, Japan
\(^3\)Department of Media Science, Nagoya University, Japan

wang@sys.eng.shizuoka.ac.jp, nakagawa@slp.ics.tut.ac.jp, kitaoka@nagoya-u.jp

Abstract

We proposed a blind dereverberation method based on spectral subtraction by Multi-Channel Least Mean Square (MCLMS) algorithm for distant-talking speech recognition in our previous study [1]. In this paper, we discuss the problems of the proposed method and present some solutions. In a distant-talking environment, the length of channel impulse response is longer than the short-term spectral analysis window. By treating the late reverberation as additive noise, a noise reduction technique based on spectral subtraction was proposed to estimate power spectrum of the clean speech using power spectra of the distorted speech and the unknown impulse responses. To estimate the power spectra of the impulse responses, a Variable Step-Size Unconstrained MCLMS (VSS-UMCLMS) algorithm for identifying the impulse responses in a time domain was extended to a frequency domain. To reduce the effect of the estimation error of channel impulse response, we normalize the early reverberation by CMN instead of the spectral subtraction used by the estimated impulse response in this paper. Furthermore, our proposed method is combined with a conventional delay-and-sum beamforming. We conducted the experiments on distorted speech signal simulated by convolving multi-channel impulse responses with clean speech. The modified proposed method achieved a relative error reduction rate of 22.7% from conventional CMN and 12.0% from the original proposed method, respectively. By combining the modified proposed method with the beamforming, a furthermore improvement (relative error reduction rate of 23.3%) was achieved.

Index Terms: distant-talking speech recognition, blind dereverberation, Multi-channel LMS, spectral subtraction, CMN.

1. Introduction

Hands-free speech recognition has been more and more popular in some special environments such as an office or a cabin of a car. Unfortunately, in a distant-talking environment, channel distortion may drastically degrade speech recognition performance.

Compensating an input feature is the main way to reduce a mismatch between the practical environment and the training environment. Cepstral Mean Normalization (CMN) has been used to reduce channel distortion as a simple and effective way of normalizing the feature space [2]. In order to be effective for CMN, the length of the channel impulse response needs to be shorter than the short-term spectral analysis window. However, the duration of the impulse response of reverberation usually has a much longer tail in a distant-talking environment. Therefore, the conventional CMN is not effective under these conditions. Several studies have focused on decreasing the above problem. A reverberation compensation method for speaker recognition using spectral subtraction in which the late reverberation was treated as additive noise was proposed in [3]. However, the drawback of this approach is that the optimum parameters for spectrum subtraction are empirically estimated on a development dataset and the late reverberation cannot be subtracted well since the late reverberation is not modelled precisely. In [4, 5], a novel dereverberation method utilizing multi-step forward linear prediction was proposed. They estimated the linear prediction coefficients in a time domain and suppress amplitude of late reflections using spectral subtraction in a spectral domain.

In our previous study [1], we proposed a blind reverberation reduction method based on spectral subtraction by adaptive Multi-Channel Least Mean Square (MCLMS) algorithm for distant-talking speech recognition. Speech captured by distant-talking microphones is distorted by the reverberation. With long impulse response, the spectrum of the distorted speech is approximated by convolving the spectrum of clean speech with the spectrum of impulse response. We treat the late reverberation as additive noise, and a noise reduction technique based on spectral subtraction can be easily applied to compensate for the late reverberation. The compensation parameter (that is, the spectrum of the impulse response) for spectral subtraction is required. In [7, 8], an adaptive MCLMS algorithm was proposed to blindly identify the channel impulse response in a time domain. In [1], we extended this method to blindly estimate the spectrum of impulse response for the spectral subtraction in a frequency domain.

The estimation error of channel impulse response is inevitable, which results in unreliable estimation of power spectrum of clean speech. On the other hand, CMN is robust to reduce the channel distortion within the spectral analysis window. In this paper, the early reverberation is normalized by CMN, and then the late reverberation is normalized by the proposed reverberation compensation technique based on the spectral subtraction by multi-channel LMS algorithm. The proposed method relies on the assumption that there are no zeros common to all channels. However, it is known that room impulse responses have a large number of zeros close to the unit circle on the z-plane. If the channels present numerically overlapping zeros, the dereverberation performance would perform poorly. The literature [5] indicated that spatial information can be used to deal with the problem of overlapping zeros. We utilize the spatial information to deal with the same problem of overlapping zeros of our method in this paper. Furthermore, a delay-and-sum beamforming is applied to the multi-channel speech compensated by the proposed method.
2. Dereverberation Based on Spectral Subtraction

When speech signal \( s[t] \) is corrupted by convolutional noise \( h[t] \) and additive noise \( n[t] \), the observed speech \( x[t] \) becomes

\[
x[t] = h[t] \otimes s[t] + n[t].
\]

(1)

In this paper, additive noise is ignored for simplification, so Eq. (1) becomes \( x[t] = h[t] \otimes s[t] \).

To analyze the effect of impulse response, the impulse response \( h[t] \) can be separated into two parts \( h_{\text{early}}[t] \) and \( h_{\text{late}}[t] \) as [3]

\[
h_{\text{early}}[t] = \begin{cases} h[t] & t < T \\ 0 & \text{otherwise} \end{cases}, h_{\text{late}}[t] = \begin{cases} h[t + T] & t \geq 0 \\ 0 & \text{otherwise} \end{cases},
\]

(2)

where \( T \) is the length of the spectral analysis window, and \( h_{\text{early}}[t] = h_{\text{late}}[t] = \delta(t - T) \otimes h_{\text{late}}[t] \). \( \delta() \) is a dirac delta function (that is, a unit impulse function). The formula (1) can be rewritten as

\[
x[t] = s[t] \otimes h_{\text{early}}[t] + s[t] \otimes h_{\text{late}}[t],
\]

(3)

where the early effect is within a frame (analysis window), and the late effect is over multiple frames.

When the length of impulse response is much shorter than analysis window size \( T \) used for short-time Fourier transform (STFT), STFT of distorted speech equals STFT of clean speech multiplied by STFT of impulse response \( h[t] \) (in this case, \( h[t] = h_{\text{early}}[t] \)). However, when the length of impulse response is much longer than an analysis window size, STFT of distorted speech is usually approximated by

\[
X(t, \omega) \approx S(t, \omega) \otimes H(\omega)
\]

\[
= S(t, \omega)H(0, \omega) + \sum_{d=1}^{D-1} S(t - d, \omega)H(d, \omega),
\]

(4)

where \( H(d, \omega) \) denotes the part of \( H(\omega) \) corresponding to frame delay \( d \). That is to say, with long impulse response, the channel distortion is no more of multiplicative nature in a linear spectral domain, rather it is convolutional.

In [3], the early term of Eq. (3) was compensated by the conventional CMN, whereas the late term of Eq. (3) was treated as additive noise, and a noise reduction technique based on spectrum subtraction was applied as

\[
\hat{S}(t, \omega) = \max(X(t, \omega) - \alpha \cdot g(\omega)X(t - T, \omega), \beta \cdot X(t, \omega)),
\]

(5)

where \( \alpha \) is the noise overestimation factor, and \( \beta \) is the spectral floor parameter to avoid negative or underflow values. However, the drawback of this approach is that the optimum parameters \( \alpha, \beta, \) and \( g(\omega) \) for the spectrum subtraction is empirically estimated on a development dataset and the STFT of late effect of impulse response as the second term of the right-hand side of Eq. (4) is not straightforward subtracted since the late reverberation is not modelled precisely.

In our previous study [1], we propose a dereverberation method based on spectral subtraction to estimate the STFT of the clean speech \( S(t, \omega) \) based on Eq. (4), and the spectrum of the impulse response for the spectral subtraction is blindly estimated using the method described in Section 3. Assuming that phases of different frames is noncorrelated for simplification, the power spectrum of Eq. (4) can be approximated as

\[
|X(t, \omega)|^2 \approx |S(t, \omega)|^2|H(0, \omega)|^2 + \sum_{d=1}^{D-1} |S(t - d, \omega)|^2|H(d, \omega)|^2,
\]

(6)

The power spectrum of clean speech \( |S(t, \omega)|^2 \) can be estimated as

\[
|\hat{S}(t, \omega)|^2 = \max(|X(t, \omega)|^2 - \alpha \sum_{d=1}^{D-1} |S(t - d, \omega)|^2|H(d, \omega)|^2, \beta \cdot |X(t, \omega)|^2)/|H(0, \omega)|^2,
\]

(7)

where \( H(d, \omega), d = 0, 1, \ldots, D - 1 \) is the STFT of impulse response which can be calculated from known impulse response or can be blindly estimated.

3. Compensation Parameter Estimation for Spectral Subtraction by Multi-channel LMS Algorithm

In [7, 8], an adaptive multi-channel LMS algorithm for blind Single-Input Multiple-Output (SIMO) system identification was proposed.

Before introducing the MCLMS algorithm for the blind channel identification, we express what SIMO systems are blind identifiable. According to [9], the following two assumptions are made to guarantee an identifiable system:

1. The polynomials formed from \( h_{\alpha}, n = 1, 2, \ldots, N \) where \( h_{\alpha} \) is \( n \)-th impulse response and \( N \) is the channel number, are co-prime, i.e., the channel transfer functions \( H(\omega) \) do not share any common zeros;

2. The autocorrelation matrix \( R_{ss} = E\{s(t)s^T(t)\} \) of input signal is of full rank (such that the single-input multiple-output (SIMO) system can be fully excited).

In the absence of additive noise, we can take advantage of the fact that

\[
x_i \ast h_j = s \ast h_i \ast h_j = x_j \ast h_i, i, j = 1, 2, \ldots, N, i \neq j,
\]

(8)

and have the following relation at time \( t \):

\[
x_i^T(t)h_j(t) = x_j^T(t)h_i(t), i, j = 1, 2, \ldots, N, i \neq j,
\]

(9)

\[
h_{\alpha}(t) = [h_{\alpha}(t, 0) \ h_{\alpha}(t, 1) \ \ldots \ h_{\alpha}(t, L - 1)]^T
\]

(10)

where \( h_{\alpha}(t) \) is \( n \)-th impulse response corresponding to \( n \)-th microphone at time \( t \) and \( h_{\alpha}(t, L) \) is \( L \)-th tap of \( n \)-th impulse response at time \( t \) and

\[
x_{\alpha}(t) = [x_{\alpha}(t) x_{\alpha}(t - 1) \ \ldots \ x_{\alpha}(t - L + 1)]^T, \ n = 1, 2, \ldots, N,
\]

(11)

where \( x_{\alpha}(t) \) is speech signal received from \( n \)-th channel at time \( t \) and \( L \) is the number of taps of the impulse response.

When the estimation of channel impulse responses deviates from the true value, an error vector is produced:

\[
e_i(t+1) = \tilde{x}_i^T(t+1)h_i(t) - \tilde{x}_j^T(t)h_j(t), i, j = 1, 2, \ldots, N, i \neq j.
\]

(12)

This error can be used to define a cost function as

\[
J(t + 1) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e_{ij}(t+1)
\]

(13)

By minimizing the cost function \( J(t + 1) \) of Eq. (13), impulse response is blindly derived. There are various methods to minimize the cost function \( J(t + 1) \), for example, constrained MCLMS algorithm, constrained Multi-Channel Newton (MCN) algorithm and Variable Step-Size Unconstrained
Table 1: Detail record conditions for impulse responses measurement. “angle”: recorded direction between microphone and loudspeaker. “RT60 (second)”: reverberation time in room. “S”: small, “L”: large.

<table>
<thead>
<tr>
<th>array no</th>
<th>array type</th>
<th>room</th>
<th>angle</th>
<th>RT60</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>linear</td>
<td>tatami-floored room (S)</td>
<td>120°</td>
<td>0.47</td>
</tr>
<tr>
<td>2</td>
<td>circle</td>
<td>tatami-floored room (S)</td>
<td>120°</td>
<td>0.47</td>
</tr>
<tr>
<td>3</td>
<td>circle</td>
<td>tatami-floored room (L)</td>
<td>130°</td>
<td>0.60</td>
</tr>
<tr>
<td>4</td>
<td>circle</td>
<td>tatami-floored room (L)</td>
<td>90°</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>linear</td>
<td>Conference room</td>
<td>50°</td>
<td>0.78</td>
</tr>
<tr>
<td>6</td>
<td>linear</td>
<td>echo room (panel)</td>
<td>70°</td>
<td>1.30</td>
</tr>
</tbody>
</table>

(VSS-UMCLMS) algorithm and so forth [7, 8]. Among these methods, the VSS-UMCLMS achieves a nice balance between complexity and convergence speed [8]. Moreover, the VSS-UMCLMS is more practical and much easier to use since the step size does not have to be specified in advance. Therefore, in this paper, we apply VSS-UMCLMS algorithm to identify the multi-channel impulse responses. The details of the VSS-UMCLMS were described in [8].

In [1], we extended the VSS-UMCLMS algorithm in a time domain to a frequency domain to estimate the spectrum of the impulse response.

4. Combining Spectral Subtraction with CMN

The estimated power spectrum of clean speech may not be very accurate due to the estimation error of the impulse response, especially the estimation error of early part of the impulse response. In addition, the unreliable estimated power spectrum of clean speech in a previous time (frame) causes a furthermore estimation error in the current time (frame). In this paper, we compensate the early reverberation by subtracting the cepstral mean of the utterance and then compensate the late reverberation by the proposed reverberation compensation method.

As is well known, cepstrum of the input speech x(t) is calculated as:

\[ C_x = \text{IDFT}(\log(|X(\omega)|^2)) \]  

where \( X(\omega) \) is the spectrum of the input speech x(t).

The early reverberation is normalized by the cepstral mean \( \bar{C} \) in a cepstral domain and then it is converted into a spectral domain as:

\[ |\bar{X}(\omega)|^2 = e^{\text{IDFT}(C_x - \bar{C})}. \]

Finally, the late reverberation is normalized by the spectrum subtraction method described in Sec. 2. That is, the denominator of Eq. (7) is dropped off, and it becomes:

\[ |\hat{S}(t, \omega)|^2 = \max(|X(t, \omega)|^2 - \alpha \sum_{d=1}^{D} |\hat{S}(t-d, \omega)|^2 |H(d, \omega)|^2, \beta \cdot |X(t, \omega)|^2). \]

5. Experiments

5.1. Experimental setup

Multi-channel distorted speech signals simulated by convolving multi-channel impulse responses with clean speech were used to evaluate our proposed algorithm. Six kinds of multi-channel impulse responses measured in various acoustical reverberant environments were selected from the RWCP sound scene database [10]. A four-channel circle type or linear type microphone array was taken from a circle + linear type microphone array (30 channels). A four-channel circle type microphone array has a diameter of 30 cm, and 4 microphones are located at equal 90° intervals. Four microphones of a linear microphone array are located at 11.32 cm intervals. Impulse responses were measured at several positions which were 2 m distance from the microphone array. The sampling frequency was 48 kHz. Table 1 shows the detail record conditions for six kinds of 4 channels microphone array.

For clean speech, twenty male speakers each with a close-microphone uttered 100 isolated words. The 100 isolated words are phonetic balance common isolated words selected from Tohoku University and Panasonic isolated spoken word database [11]. The average time of all utterances was about 0.6 second. The sampling frequency was 12 kHz. The impulse responses sampled at 48 kHz were downsampling to 12 kHz to convolve with clean speech. The frame length was 21.3 ms, and the frame shift was 8 ms with a 256 point Hamming window. Then, 116 Japanese speaker-independent syllable-based HMMs (strictly speaking, mora-unit HMMs [12]) were trained using 27992 utterances read by 175 male speakers (JNAS corpus). Each continuous-density HMM had 5 states, 4 with pdfs of output probability. Each pdf consisted of 4 Gaussians with full-covariance matrices. The feature space comprised 10 MFCCs. First- and second-order derivatives of the cepstra plus first and second derivatives of the power component were also included.

The number of reverberant window \( D \) in Eq. (4) was set to 8. The length of the Hamming window for DFT was 256 (\( \approx 21.3 \text{ ms} \)), and the overlapping rate was 1/2. No special parameters such as over-subtraction parameters were used for spectral subtraction (\( \alpha = 1 \)), except that the subtracted value was controlled so that it did not become negative (\( \beta = 0.15 \)). The speech recognition performance for clean isolated words was 96.0%.

5.2. Experimental results and discussion

Table 2 shows the baseline results for speech recognition. CMN performed on distorted speech was used as baseline. The LSE based inverse filtering was an ideal condition. However, it could not appropriately deal with a non-minimum phase impulse response [13], whose case is often in real reverberant environments. Therefore, the speech recognition performance was not an upper bound even using the known impulse response. There are many other more precise inverse filtering techniques such as [13, 14] and so forth. We will use the more precise inverse filtering techniques as ideal condition in near future.

In our previous study [1], only the speech signal from the first channel of each microphone array was performed for
Table 3: Speech recognition performance of the original and modified proposed methods. 4 microphones were used to estimate the spectrum of impulse response. Delay-and-sum beamforming was performed to 4-channel dereverberant speech signals. For proposed method, each channel speech was compensated by the corresponding impulse response (%).

<table>
<thead>
<tr>
<th>distorted speech #</th>
<th>Original method [1]</th>
<th>Modified method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mic.</td>
<td>beam-forming</td>
</tr>
<tr>
<td>1</td>
<td>69.9</td>
<td>76.5</td>
</tr>
<tr>
<td>2</td>
<td>63.4</td>
<td>77.5</td>
</tr>
<tr>
<td>3</td>
<td>66.8</td>
<td>73.1</td>
</tr>
<tr>
<td>4</td>
<td>69.5</td>
<td>78.3</td>
</tr>
<tr>
<td>5</td>
<td>63.4</td>
<td>72.1</td>
</tr>
<tr>
<td>6</td>
<td>62.5</td>
<td>67.7</td>
</tr>
<tr>
<td>Ave.</td>
<td>65.9</td>
<td>74.2</td>
</tr>
</tbody>
</table>

For our original proposed method mentioned in Sec. 2, at first speech signals from 4 microphones were used to blindly identify the compensation parameters for the spectral subtraction (that is, the spectra of the channel impulse responses), and then the spectrum of the first channel impulse response was used to compensate for the reverberation of the speech signal from the first channel. In this paper, the modified proposed method described in Sec. 4 was also evaluated. Moreover, the delay-and-sum beamforming was performed to the multi-channel dereverberate speech for both the original and modified proposed methods.

Table 3 shows the experimental results of the original and modified proposed methods for speech recognition. For our proposed methods, CMN was also performed on the dereverberant speech. The original proposed method [1] remarkably improved speech recognition performance. However, for the #2 of the distorted speech, the performance of dereverberation speech based on the original method was worse than that of the CMN. The modified proposed method improved the speech recognition significantly than the original method, and it was better than the CMN for all severe reverberant conditions. When the delay-and-sum beamforming was combined to our proposed method, a further improvement was achieved. A relative error reduction rate of 23.3% from the original proposed method for a single microphone, 23.4% from the modified method for a single microphone were achieved, respectively. Comparing the conventional CMN combining with beamforming (69.4% in Table 2), a relative error reduction rate of 24.8% was achieved. It outperformed the result through inverse filtering.

### 6. Conclusions and Future Work

In this paper, we improved a blind reverberation reduction method based on spectral subtraction by MCLMS algorithm for distant-talking speech recognition. In a distant-talking environment, the length of channel impulse response is longer than the short-term spectral analysis window. Therefore, the channel distortion is not even with multiplicative nature in a linear spectral domain, rather it is convolutional. We treated the late reverberation as additive noise, and a noise reduction technique based on spectrum subtraction was proposed to estimate the clean power spectrum. Power spectrum of impulse response was necessary to estimate the clean power spectrum [1]. To estimate the power spectra of the impulse responses, a VSS-UMCLMS algorithm for identifying the impulse responses in a time domain was extended to the frequency domain [1]. The estimation error of channel impulse response is inevitable, which results in unreliable estimation of power spectrum of clean speech. In this paper, the early reverberation was normalized by CMN, and then the late reverberation was normalized by the proposed spectral subtraction by multi-channel LMS algorithm. A delay-and-sum beamforming was also applied to the multi-channel speech compensated by the proposed reverberation compensation technique based on the spectral subtraction. Our original and modified proposed algorithms were evaluated by distorted speech signals simulated by convolving multi-channel impulse responses with clean speech taken from Tohoku University and Panasonic isolated spoken word database. The modified proposed method achieved an average relative error reduction rate of 22.7% from the conventional CMN and 12.0% from the original proposed method, respectively. By combining the modified proposed method with beamforming, a furthermore improvement (relative error reduction rate of 23.3%) was achieved. We try to evaluate our proposed methods by real-world speech data in the future.

### 7. References


