PAPER

Distant-Talking Speech Recognition Based on Spectral Subtraction by Multi-Channel LMS Algorithm

Longbiao WANG[†a)], **Norihide KITAOKA**[††], ***Members***, ***and*** **Seiichi NAKAGAWA**[†††], ***Fellow***

SUMMARY We propose a blind dereverberation method based on spectral subtraction using a multi-channel least mean squares (MCLMS) algorithm for distant-talking speech recognition. In a distant-talking environment, the 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 the 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 is extended to a frequency domain. To reduce the effect of the estimation error of the channel impulse response, we normalize the early reverberation by cepstral mean normalization (CMN) instead of spectral subtraction using the estimated impulse response. Furthermore, our proposed method is combined with conventional delay-and-sum beamforming. We conducted recognition experiments on a distorted speech signal simulated by convolving multi-channel impulse responses with clean speech. The proposed method achieved a relative error reduction rate of 22.4% in relation to conventional CMN. By combining the proposed method with beamforming, a relative error reduction rate of 24.5% in relation to the conventional CMN with beamforming was achieved using only an isolated word (with duration of about 0.6 s) to estimate the spectrum of the impulse response.

key words: *distant-talking speech recognition, blind dereverberation, multi-channel least mean squares, spectral subtraction, cepstral mean normalization*

1. Introduction

In a distant-talking environment, channel distortion drastically degrades speech recognition performance. Compensating an input feature is the main way to reduce mismatch between the practical environment and the training environment. Cepstral mean normalization (CMN) has been employed to reduce channel distortion as a simple and effective way of normalizing the feature space [1], [2]. To be suitable for CMN, the length of the channel impulse response needs to be shorter than the short-term spectral analysis window. However, the impulse response of reverberation usually has a much longer tail in a distant-talking environment. Therefore, conventional CMN is not sufficiently effective under these conditions. Several studies have focused on mitigating the above problem. Raut et al. [3], [4] used preceding states as units of preceding speech segments, and they adapted models accordingly by estimating their contributions to the current state using a maximum likelihood function. However, model adaptation using *a priori* training data makes the models less practical to use because the true impulse response or matched reverberant utterance is not always expected for various environments. A blind deconvolution-based approach for the restoration of speech degraded by the acoustic environment was proposed in [5]. The proposed scheme processed the outputs of two microphones using cepstra operations and the theory of signal reconstruction from the phase only. Avendano et al. explored a speech dereverberation technique whose principle was the recovery of the envelope modulations of the original (anechoic) speech [6], [7]. They applied a technique that they originally developed to treat background noise [8] to the dereverberation problem. A novel approach for multimicrophone speech dereverberation was proposed in [9]. The method was based on the construction of the null subspace of the data matrix in the presence of colored noise, employing generalized singular-value decomposition or generalized eigenvalue decomposition of the respective correlation matrices. A reverberation compensation method for speaker recognition using spectral subtraction in which the late reverberation is treated as additive noise was proposed in [10], [11]. However, the drawback of this approach is that the optimum parameters for spectral subtraction are empirically estimated from a development dataset and the late reverberation cannot be subtracted well since it is not modeled precisely. In [12], [13], a novel dereverberation method utilizing multi-step forward linear prediction was proposed. They estimated the linear prediction coefficients in a time domain and suppressed the amplitude of late reflections through spectral subtraction in a spectral domain.

In this paper, we propose a robust distant-talking speech recognition method based on spectral subtraction employing the adaptive multi-channel least mean squares (MCLMS) algorithm. Speech captured by distant-talking microphones is distorted by the reverberation. With a long impulse response, the spectrum of the distorted speech is approximated by convolving the spectrum of clean speech with the spectrum of the impulse response as explained in the next section. This enables us to 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. By excluding the phase informa-

Manuscript received June 15, 2010.
Manuscript revised October 23, 2010.

[†]The author is with Shizuoka University, Hamamatsu-shi, 432–8561 Japan.
[††]The author is with Nagoya University, Nagoya-shi, 465–8603 Japan.
[†††]The author is with Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
a) E-mail: wang@sys.eng.shizuoka.ac.jp

DOI: 10.1587/transinf.E94.D.659

tion from the dereverberation operation as in [13], [14], the dereverberation reduction in a power spectral domain provides robustness against certain errors that the conventional sensitive inverse filtering method cannot achieve. The compensation parameter (that is, the spectrum of the impulse response) for spectral subtraction is required. In [15]–[18], an adaptive MCLMS algorithm was proposed to blindly identify the channel impulse response in a time domain. In this paper, we extend the method to blindly estimate the spectrum of the impulse response for spectral subtraction in a frequency domain.

Error in estimating the channel impulse response is inevitable and results in unreliable estimation of the power spectrum of clean speech. On the other hand, CMN is robust such that it reduces channel distortion within the spectral analysis window. In this paper, the early reverberation is normalized by CMN, and the late reverberation is then normalized by the proposed reverberation compensation technique based on spectral subtraction by a multi-channel LMS algorithm. Therefore, another novel point of our proposed method is that it replaces the compensation of early reverberation with CMN with rigorous derivation. Furthermore, delay-and-sum beamforming is applied to the multi-channel speech compensated by the proposed method.

The remainder of this paper is organized as follows. Section 2 describes our proposed dereverberation method based on spectral subtraction. A multi-channel method based on the LMS algorithm and used to estimate the power spectrum of the impulse response (that is, a compensation parameter for spectral subtraction) is briefly described in Sect. 3. In Sect. 4, we normalize the early reverberation by CMN instead of spectral subtraction using the estimated impulse response. Section 5 describes the experimental results of distant-talking speech recognition using multi-channel distorted speech signals simulated by convolving multi-channel impulse responses with clean speech. Finally, Sect. 6 summarizes the paper and describes future directions of research.

2. Dereverberation Based on Spectral Subtraction

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

$$x[t] = h[t] * s[t] + n[t]. \tag{1}$$

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

To analyze the effect of impulse response, the impulse response $h[t]$ can be separated into two parts $h_{early}[t]$ and $h_{late}[t]$ as [10], [11]

$$h_{early}[t] = \begin{cases} h[t] & \text{t} < \text{T} \\ 0 & \text{otherwise} \end{cases},$$
$$h_{late}[t] = \begin{cases} h[t+T] & \text{t} \geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

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

$$x[t] = s[t] * h_{early}[t] + s[t-T] * h_{late}[t], \tag{3}$$

where the early effect is distortion within a frame (analysis window), and the late effect comes from previous 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, $h[t] = h_{early}[t]$). However, when the length of impulse response is much longer than an analysis window size, STFT of distorted speech is usually approximated by

$$\begin{aligned} X(f,\omega) &\approx S(f,\omega) * H(\omega) \\ &= S(f,\omega)H(0,\omega) + \sum_{d=1}^{D-1} S(f-d,\omega)H(d,\omega), \end{aligned} \tag{4}$$

where f is frame index, $H(\omega)$ is STFT of impulse response, $S(f,\omega)$ is STFT of clean speech s and $H(d,\omega)$ 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 [4].

In [11], 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 spectral subtraction was applied as

$$|\hat{S}(f,\omega)| = \max(|X(f,\omega)| - \alpha \cdot g(\omega)|X(f-1,\omega)|, \beta \cdot |X(f,\omega)|), \tag{5}$$

where α is the noise overestimation factor, β is the spectral floor parameter to avoid negative or underflow values, and $g(\omega)$ is a frequency-dependent value which is determined on a development and set as $|1 - 0.9e^{j\omega}|$ [11]. However, the drawback of this approach is that the optimum parameters α, β, and for the spectral 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 this paper, we propose a dereverberation method based on spectral subtraction to estimate the STFT of the clean speech $\hat{S}(f,\omega)$ based on Eq. (4), and the spectrum of the impulse response for the spectral subtraction is blindly estimated using the method described in Sect. 3. Assuming that phases of different frames is noncorrelated for simplification, the power spectrum of Eq. (4) can be approximated as

$$\begin{aligned} |X(f,\omega)|^2 &\approx |S(f,\omega)|^2 |H(0,\omega)|^2 \\ &\quad + \sum_{d=1}^{D-1} |S(f-d,\omega)|^2 |H(d,\omega)|^2. \end{aligned} \tag{6}$$

$$|\hat{S}(f,\omega)|^2 = \frac{\max(|X(f,\omega)|^2 - \alpha \cdot \sum_{d=1}^{D-1} |\hat{S}(f-d,\omega)|^2 |H(d,\omega)|^2, \beta \cdot |X(f,\omega)|^2)}{|H(0,\omega)|^2}, \quad (7)$$

The power spectrum of clean speech $|\hat{S}(f,\omega)|^2$ can be estimated as Eq. (7) (see the next page), 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

3.1 Blind Channel Identification in Time Domain

3.1.1 Identifiability and Principle

In [15]–[18], 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*. A multi-channel FIR (Finite Impulse Response) system can be blindly primarily because of the channel diversity. As an extreme counter-example, if all channels of a SIMO system are identical, the system reduces to a Single-Input Single-Output (SISO) system, becoming unidentifiable. In addition, the source signal needs to have sufficient modes to make the channels fully excited. According to [19], the following two assumptions are made to guarantee an identifiable system:

1. The polynomials formed from $h_n, n = 1, 2, \cdots, N$, where h_n is *n-th* impulse response and N is the channel number, are co-prime [†], i.e., the channel transfer functions $H_n(z)$ do not share any common zeros;
2. The autocorrelation matrix $\mathbf{R}_{ss} = E\{s(k)s^T(k)\}$ of input signal is of full rank (such that the single-input multiple-output (SIMO) system can be fully excited).

In the following, these two conditions are assumed to hold so that we will be dealing with a blindly identifiable FIR (Finite Impulse Response) SIMO system.

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

$$x_i * h_j = s * h_i * h_j = x_j * h_i,\ i, j = 1, 2, \cdots, N, i \neq j, \quad (8)$$

and have the following relation at time t:

$$\mathbf{x}_i^T(t)\mathbf{h}_j(t) = \mathbf{x}_j^T(t)\mathbf{h}_i(t), \quad i, j = 1, 2, \cdots, N, i \neq j, \quad (9)$$

where $\mathbf{h}_i(t)$ is the *i-th* impulse response at time t and

$$\mathbf{x}_n(t) = [x_n(t)\ \ x_n(t-1)\ \ \cdots\ \ x_n(t-L+1)]^T, \quad n = 1, 2, \cdots, N, \quad (10)$$

where $\mathbf{x}_n(t)$ is speech signal received from the *n-th* channel at time t and L is the number of taps of the impulse response. Multiplying Eq. (9) by $\mathbf{x}_n(t)$ and taking expectation yields,

$$\mathbf{R}_{x_i x_i}(t+1)\mathbf{h}_j(t) = \mathbf{R}_{x_i x_j}(t+1)\mathbf{h}_i(t), \quad i, j = 1, 2, \cdots, N, i \neq j, \quad (11)$$

where $\mathbf{R}_{x_i x_j}(t+1) = E\{\mathbf{x}_i(t+1)\mathbf{x}_j^T(t+1)\}$. Eq. (11) comprises $N(N-1)$ distinct equations. By summing up the $N-1$ cross relations associated with one particuar channel $\mathbf{h}_j(t)$, we get

$$\sum_{i=1, i\neq j}^{N} \mathbf{R}_{x_i x_i}(t+1)\mathbf{h}_j(t) = \sum_{i=1, i\neq j}^{N} \mathbf{R}_{x_i x_j}(t+1)\mathbf{h}_i(t), \quad j = 1, 2, \cdots, N. \quad (12)$$

Over all channels, we then have a total of N equations. In matrix form, this set of equations is written as:

$$\mathbf{R}_{x+}(t+1)\mathbf{h}(t) = 0, \quad (13)$$

where

$$\mathbf{h}(t) = [\mathbf{h}_1(t)^T\ \ \mathbf{h}_2(t)^T\ \ \cdots\ \ \mathbf{h}_N(t)^T]^T, \quad (15)$$

$$\mathbf{h}_n(t) = [h_n(t,0)\ \ h_n(t,1)\ \ \cdots\ \ h_n(t,L-1)]^T, \quad (16)$$

where $h_n(t,l)$ is the *l-th* tap of the *n-th* impulse response at time t. If the SIMO system is blindly identifiable, the matrix $\mathbf{R}_{x+}$ is rank deficient by 1 (in the absence of noise) and the channel impulse responses can be uniquely determined.

When the estimation of channel impulse responses is deviated from the true value, an error vector at time $t+1$ is produced by:

$$\mathbf{e}(t+1) = \tilde{\mathbf{R}}_{x+}(t+1)\hat{\mathbf{h}}(t), \quad (17)$$

where $\tilde{\mathbf{R}}_{x_i x_j}(t+1) = \mathbf{x}_i(t+1)\mathbf{x}_j^T(t+1), i, j = 1, 2, \cdots, N$ and $\hat{\mathbf{h}}(t)$ is the estimated model filter at time t. Here we put a tilde in $\tilde{\mathbf{R}}_{x_i x_j}$ to distinguish this instantaneous value from its mathematical expectation $\mathbf{R}_{x_i x_j}$.

This error can be used to define a cost function at time $t+1$

$$J(t+1) = \|\mathbf{e}(t+1)\|^2 = \mathbf{e}(t+1)^T\mathbf{e}(t+1). \quad (19)$$

By minimizing the cost function J of Eq. (19), the impulse response is blindly derived. There are various methods to minimize the cost function J, for example, constrained Multi-Channel LMS (MCLMS) algorithm, constrained Multi-Channel Newton (MCN) algorithm and Variable Step-Size Unconstrained MCLMS (VSS-UMCLMS) algorithm and so forth [16], [18]. Among these methods,

[†] In mathematics, the integers a and b are said to be co-prime if they have no common factor other than 1, or equivalently, if their greatest common divisor is 1.

$$\mathbf{R}_{x+}(t+1)=\begin{bmatrix} \sum_{n\neq 1}\mathbf{R}_{x_n x_n}(t+1) & -\mathbf{R}_{x_2 x_1}(t+1) & \cdots & -\mathbf{R}_{x_N x_1}(t+1) \\ -\mathbf{R}_{x_1 x_2}(t+1) & \sum_{n\neq 2}\mathbf{R}_{x_n x_n}(t+1) & \cdots & -\mathbf{R}_{x_N x_2}(t+1) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{x_1 x_N}(t+1) & -\mathbf{R}_{x_2 x_N}(t+1) & \cdots & \sum_{n\neq N}\mathbf{R}_{x_n x_n}(t+1) \end{bmatrix}, \tag{14}$$

$$\tilde{\mathbf{R}}_{x+}(t+1)=\begin{bmatrix} \sum_{n\neq 1}\tilde{\mathbf{R}}_{x_n x_n}(t+1) & -\tilde{\mathbf{R}}_{x_2 x_1}(t+1) & \cdots & -\tilde{\mathbf{R}}_{x_N x_1}(t+1) \\ -\tilde{\mathbf{R}}_{x_1 x_2}(t+1) & \sum_{n\neq 2}\tilde{\mathbf{R}}_{x_n x_n}(t+1) & \cdots & -\tilde{\mathbf{R}}_{x_N x_2}(t+1) \\ \vdots & \vdots & \ddots & \vdots \\ -\tilde{\mathbf{R}}_{x_1 x_N}(t+1) & -\tilde{\mathbf{R}}_{x_2 x_N}(t+1) & \cdots & \sum_{n\neq N}\tilde{\mathbf{R}}_{x_n x_n}(t+1) \end{bmatrix}, \tag{18}$$

the VSS-UMCLMS achieves a nice balance between complexity and convergence speed [18]. 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.

3.1.2 Variable Step-Size Unconstrained Multi-Channel LMS Algorithm in Time Domain

The cost function $J(t+1)$ at time $t+1$ diminishes and its gradient with respect to $\hat{\mathbf{h}}(t)$ can be approximated as

$$\Delta J(t+1)\approx\frac{2\tilde{\mathbf{R}}_{x+}(t+1)\hat{\mathbf{h}}(t)}{\|\hat{\mathbf{h}}(t)\|^2} \tag{20}$$

and the model filter $\hat{\mathbf{h}}(t+1)$ at time $t+1$ is

$$\hat{\mathbf{h}}(t+1)=\hat{\mathbf{h}}(t)-2\mu\tilde{\mathbf{R}}_{x+}(t+1)\hat{\mathbf{h}}(t), \tag{21}$$

which is theoretically equivalent to the adaptive algorithm proposed in [20] although the cost functions are defined in different ways in these two adaptive blind SIMO identification algothrithms. In Eq. (21), μ is step size for Multi-channel LMS.

With such a simplified adaptive algorithm, the primary concern is whether it would converge to the trivial all-zero estimate. Fortunately this will not happen as long as the initial estimate $\hat{\mathbf{h}}(0)$ is not orthogonal to the true channel impulse response vector $\mathbf{h}$, as shown in [20].

Finally, an optimal step size for the unconstrained MCLMS at time $t+1$ is obtained by

$$\mu_{opt}(t+1)=\frac{\hat{\mathbf{h}}^T(t)\Delta J(t+1)}{\|\Delta J(t+1)\|^2}. \tag{22}$$

The details of the VSS-UMCLMS were described in [18].

3.2 Extending VSS-UMCLMS Algorithm to Compensation Parameter Estimation for Spectral Subtraction

To blindly estimate the compensation parameter (that is, the spectrum of impulse response), we extend the MCLMS algorithm mentioned in Sect. 3.1 from a time domain to a frequency domain in this section.

The spectrum of distorted signal is a convolution operation of the spectrum of clean speech and that of impulse response as shown in Eq. (4). The spectrum of the impulse response is dependent on frequency ω, and the varibale ω is omitted for simplification. Thus, in the absence of additive noise, the spectra of distorted signals have the following relation at frame f on the frequency domain:

$$\mathbf{X}_i^T(f)\mathbf{H}_j(f)=\mathbf{X}_j^T(f)\mathbf{H}_i(f),\quad i,j=1,2,\ldots,N,\quad i\neq j, \tag{23}$$

Where $\mathbf{X}_n(f)=[X_n(f)\quad X_n(f-1)\quad\ldots\quad X_n(f-D+1)]^T$ is a D-dimention vector of spectra of the distorted speech received from the *n-th* channel at frame f, $X_n(f)$ is the spectrum of the distorted speech received from the *n-th* channel at frame f for frequency ω, $\mathbf{H}_n(f)=[H_n(f,0)\ H_n(f,1)\ \ldots\ H_n(f,d)\ \ldots\ H_n(f,D-1)]^T, d=0,1,\ldots,D-1$ is a D-dimensional vector of spectra of the impulse response, and $H_n(f,d)$ is the spectrum of the impulse response for frequency ω at frame f corresponding to frame delay d (that is, at frame $f+d$).

Using Eq. (23) in place of Eq. (9), the spectra of the impulse responses can be blindly estimated by the VSS-UMCLMS mentioned in Sect. 3.1.2.

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 frame causes a furthermore estimation error in the current 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:

$$C_x = IDFT(\log(|X(\omega)|^2)) \tag{24}$$

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 (linear cepstrum is used) and then it is converted into a spectral domain as:

$$|\tilde{X}(\omega)|^2 = |e^{DFT(C_x - \bar{C})}| = \frac{|X(f,\omega)|^2}{|\bar{X}(f,\omega)|^2}, \tag{25}$$

where $\bar{X}(f,\omega)$ is mean vector of $X(f,\omega)$. After this normalization processing, the Eq. (6) becomes as

$$\begin{aligned}
&|\tilde{X}(f,\omega)|^2 \\
&= \frac{|X(f,\omega)|^2}{|\bar{X}(f,\omega)|^2} \\
&= \frac{|S(f,\omega)|^2|H(0,\omega)|^2}{|\bar{X}(f,\omega)|^2} + \sum_{d=1}^{D-1}\{\frac{|S(f-d,\omega)|^2|H(d,\omega)|^2}{|\bar{X}(f,\omega)|^2}\} \\
&\approx \frac{|S(f,\omega)|^2}{|\bar{S}(f,\omega)|^2} + \sum_{d=1}^{D-1}\{\frac{|S(f-d,\omega)|^2}{|\bar{S}(f,\omega)|^2} \times \frac{|H(d,\omega)|^2}{|H(0,\omega)|^2}\} \\
&= |\tilde{S}(f,\omega)|^2 + \frac{\sum_{d=1}^{D-1}\{|\tilde{S}(f-d,\omega)|^2 \times |H(d,\omega)|^2\}}{|H(0,\omega)|^2},
\end{aligned} \tag{26}$$

where $|\tilde{S}(f,\omega)|^2 = \frac{|S(f,\omega)|^2}{|\bar{S}(f,\omega)|^2}$, $|\bar{X}(f,\omega)|^2 \approx |\bar{S}(f,\omega)|^2 \times |H(0,\omega)|^2$, and $\bar{S}(f,\omega)$ is mean vector of $S(f,\omega)$. The estimated clean power spectrum $|\tilde{S}(f,\omega)|^2$ becomes as

$$|\tilde{S}(f,\omega)|^2 = |\tilde{X}(f,\omega)|^2 - \frac{\sum_{d=1}^{D-1}\{|\tilde{S}(f-d,\omega)|^2 \times |H(d,\omega)|^2\}}{|H(0,\omega)|^2}. \tag{27}$$

The spectral subtraction is used to prevent the estimated clean power spectrum being negative value, the Eq. (27) is modified as:

$$\begin{aligned}
&|\hat{S}(f,\omega)|^2 \approx \max(|\tilde{X}(f,\omega)|^2 - \\
&\alpha \cdot \frac{\sum_{d=1}^{D-1}\{|\tilde{S}(f-d,\omega)|^2|H(d,\omega)|^2\}}{|H(0,\omega)|^2}, \beta \cdot |\tilde{X}(f,\omega)|^2).
\end{aligned} \tag{28}$$

The methods given in Eq. (7) and Eq. (28) are referred to as *original proposed method* and *modified proposed method*, respectively.

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 Real World Computing Partnership sound scene database [21]. A four-channel circular or linear microphone array was taken from a circular + linear microphone array (30 channels). The four-channel circle type microphone array had a diameter of 30 *cm*, and the four microphones were located at equal 90° intervals. The four microphones of the linear microphone array were located at 11.32 *cm* intervals. Impulse responses were measured at several positions 2 *m* from the microphone array. The sampling frequency was 48 *kHz*. Table 1 details the conditions for six recordings with a four-channel microphone array.

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.

array no	array type	room	angle	RT60
1	linear	tatami-floored room (S)	120°	0.47
2	circle	tatami-floored room (S)	120°	0.47
3	circle	tatami-floored room (L)	90°	0.60
4	circle	tatami-floored room (L)	130°	0.60
5	linear	Conference room	50°	0.78
6	linear	echo room (panel)	70°	1.30

For clean speech, 20 male speakers each with a close microphone uttered 100 isolated words. The 100 isolated words were phonetically balanced common isolated words selected from the Tohoku University and Panasonic isolated spoken word database [22]. The average time of all utterances was about 0.6 s. The sampling frequency was 12 *kHz*. The impulse responses sampled at 48 *kHz* were downsampled to 12 *kHz* so that they could be convolved 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 [23]) were trained using 27,992 utterances read by 175 male speakers (JNAS corpus [24]). Each continuous-density HMM had five states, four with probability density functions (pdfs) of output probability. Each pdf consisted of four Gaussians with full-covariance matrices. The acoustic model was common for the baseline and proposed methods, and it was trained in a clean condition. The feature space comprised 10 mel-frequency cepstral coefficients. First- and second-order derivatives of the cepstra plus first and second derivatives of the power component were also included (32 feature parameters in total).

The number of reverberant windows D in Eq. (4) was set to eight, which was empirically determined. In general, the window size D is proportional to RT60. However, the window size D is also affected by the reverberation property; for example, the ratio of power of the late reverberation to the power of the early reverberation. In our preliminary experiment with partial test data, the performance of our proposed method with a window size D = 2 to 16 outperformed the baseline significantly and the window size D = 8 achieved the best result. Automatic estimation of the optimum window size D is our future work. The length of the Hamming window for discrete Fourier transformation

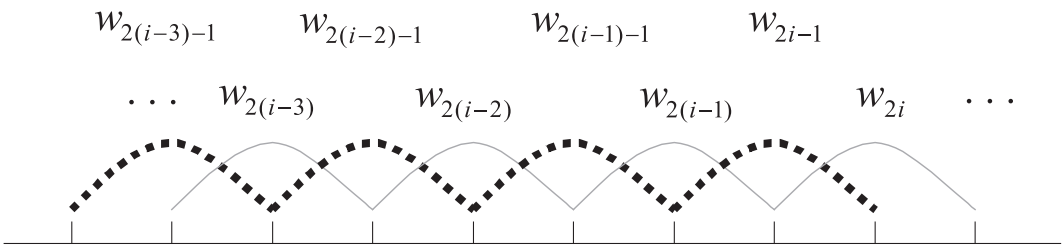


Fig. 1 Analysis window for spectral subtraction. For the proposed dereverberation based on spectral subtraction, the previous clean power spectra estimated with a skip window were used to estimate the current clean power spectrum because the frame shift is half the frame length in this paper. For example, to estimate the clean power spectrum of the *2i-th* window W_{2i}, the estimated clean power spectra of the *2(i-1)-th* window $W_{2(i-1)}$, the *2(i-2)-th* window $W_{2(i-2)}$, $\cdots$ were used.

Table 2 Baseline results (%).

distorted speech #	Single Mic. (first channel) w/o CMN	CMN	inverse filtering	SS with true impulse response	beam-forming
1	46.0	64.2	77.4	74.5	69.4
2	48.4	64.2	71.8	71.8	73.2
3	53.3	62.8	77.3	73.2	71.4
4	48.7	65.1	76.2	72.5	71.8
5	43.5	56.2	70.9	66.1	67.7
6	40.3	54.7	72.4	66.2	63.1
Ave.	46.7	61.2	74.3	70.7	69.4

was 256 (21.3 *ms*), and the rate of overlap was 1/2 [†]. An illustration of the analysis window is shown in Fig. 1. For the proposed dereverberation based on spectral subtraction, the previous clean power spectra estimated with a skip window were used to estimate the current clean power spectrum [††]. The spectrum of the impulse response $H(d, \omega)$ is estimated using the corresponding utterance to be recognized with average duration of about 0.6 second. No special parameters such as over-subtraction parameters were used in 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. "Distorted speech #" in Tables 2 and 3 corresponds to "array no" in Table 1. The CMN of the distorted speech was used as a baseline. LSE-based inverse filtering [25] using a true impulse response was the ideal condition. However, this filtering cannot appropriately deal with a non-minimum phase impulse response [25], which is common in real reverberant environments. Therefore, the speech recognition performance was not an upper bound when using the known impulse response. There are many other more precise inverse filtering techniques such as that of [25], [26]. We will use the more precise inverse filtering techniques as the ideal condition in the near future. The result of spectral subtraction with the true impulse response did not improve the performance sufficiently. The reason for this might be that optimum parameters α and β for spectral subtraction were not used and the distorted input speech was analyzed using a Hamming window while the compensation parameter $H(d, \omega)$ was calculated from the true impulse response without using a Hamming window.

In this paper, speech recognition was performed using speech data for a single microphone and multiple channels. For single-microphone processing [†††], only the speech signal from the first channel of each microphone array was used for speech recognition. In our original proposed method described in Sect. 2, speech signals from four 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 Sect. 4 is also evaluated. Moreover, delay-and-sum beamforming is performed for the multi-channel dereverberate speech in both the original and modified proposed methods.

Table 3 shows the experimental results obtained for the original and modified proposed methods for speech recognition. In our proposed methods, CMN was also performed on the dereverberant speech. In this paper, the cepstral mean was calculated using one isolated word (that is, a recognition word with duration of about 0.6 s) in both methods. The original proposed method based on Eq. (7) remarkably improved the speech recognition performance. The modified proposed method based on Eq. (28) improved speech recognition significantly compared with the original proposed method and CMN for all severe reverberant conditions. The reason was that the compensation error using spectrum sub-

[†]Our method employs two-stage processing of time-domain speech. In the first stage, the reverberant speech is transformed to a frequency domain, and a method based on spectral subtraction is used to compensate the late reverberation before the speech is transformed to the time domain. In this stage, the rate of overlap is 1/2 as shown in Fig. 1. In the second stage, the compensated time-domain speech is transformed to MFCCs with a frame shift of 8 ms.

[††]Eq. (23) is true when using a skip window and the spectrum of the impulse response can be blindly estimated.

[†††]In fact, multi-channel speech data were used to estimate the channel transfer function *a priori*.

Table 3 Speech recogniton performacnes 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 (%).

distorted speech #	Original proposed method		Modified proposed method	
	w/o beamforming	beam-forming	w/o beamforming	beam-forming
1	66.3	70.8	72.8	76.0
2	65.8	76.3	70.0	80.6
3	69.9	76.1	73.9	80.3
4	69.4	76.7	71.1	78.6
5	63.1	70.6	66.0	74.4
6	62.0	68.1	65.8	71.2
Ave.	66.1	73.1	69.9	76.9

traction was greater than that using CMN for early reverberation. The result of the proposed method was similar to that of spectral subtraction with a true impulse response. Therefore, we can state that the compensation parameter for spectral subtraction was estimated accurately. The modified proposed method [†] achieved an average relative error reduction rate of 22.4% in relation to conventional CMN and 11.2% in relation to the original proposed method. When delay-and-sum beamforming was combined with our proposed method, further improvement was achieved. Relative error reduction rates of 20.6% in relation to the original proposed method and 23.3% in relation to the modified method without beamforming were achieved. Comparing the conventional CMN combined with beamforming (69.4% in Table 2), relative error reduction rate of 24.5% were achieved.

We also analyzed the relationship between the speech recognition rate and reverberation time for conventional CMN and the proposed method. The results are shown in Fig. 2. Naturally, the speech recognition rate degraded as the reverberation time increased. Using the modified proposed method, the reduction of the speech recognition rate was less especially for an impulse response with a long reverberation time compared with conventional CMN. The reason is that our proposed method can compensate for the late reverberation through spectral subtraction using an estimated power spectrum of the impulse response.

We attempted to compare our proposed methods with the dereverberation method based on the VSS-UMCLMS algorithm in the time domain proposed in [17], [18]. However, the estimation error of the impulse response was very large. Therefore, the recognition rate of the compensated speech using the estimated impulse response was significantly worse than that of unprocessed received distorted speech. The reason might be that the tap number of the impulse response was very large and the duration of the utterance (that is, a word with duration of about 0.6 s) was very short. Therefore, the VSS-UMCLMS algorithm in the time domain might not be convergent. The other problem with the algorithm in the time domain is the estimation cost. The estimation time of the algorithm in the time domain was about 360 times that in the frequency domain under the experimental setup described in Sect. 5.1.

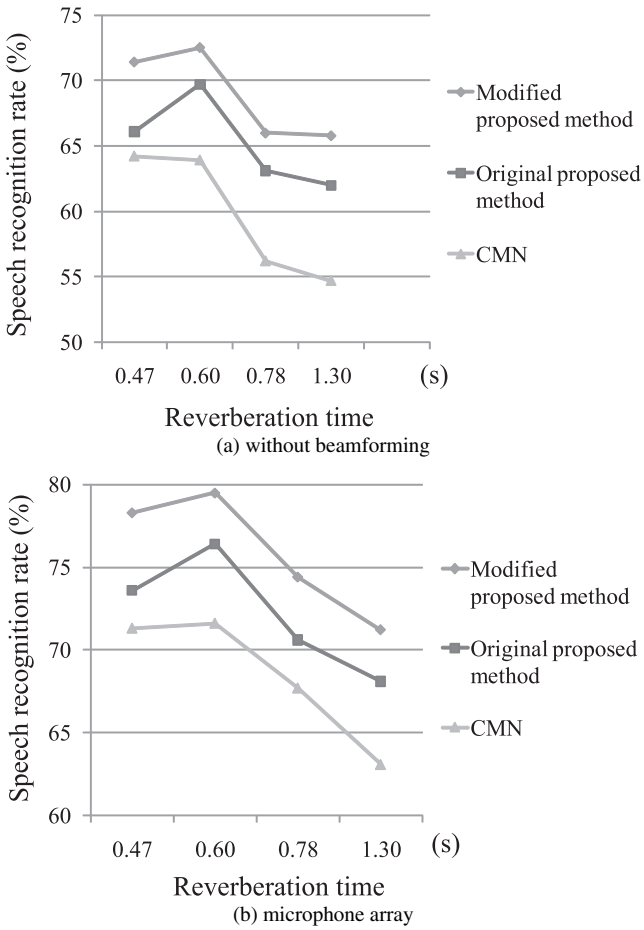


Fig. 2 The relationship between speech recognition rate and reverberation time.

6. Conclusions and Future Work

In this paper, we proposed a blind reverberation reduction method based on spectral subtraction employing a variable step-size unconstrained multi-channel LMS algorithm for distant-talking speech recognition. In a distant-talking environment, the channel distortion no longer has a 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 spectral subtraction was proposed to estimate the clean power spectrum. The power spectrum of the impulse response was required to estimate the clean power spectrum. 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. Error in estimating the channel impulse response is inevitable and results in unreliable estimation of the power spectrum of clean speech. In this paper, the early reverberation was normalized by

[†]Our proposed method could not achieve perfect speech recognition because of the error in estimating the impulse response.

CMN, and then the late reverberation was normalized by the proposed spectral subtraction employing a multi-channel LMS algorithm. Delay-and-sum beamforming was also applied to the multi-channel speech compensated by the proposed reverberation compensation technique based on spectral subtraction. Our original and modified proposed algorithms were evaluated using distorted speech signals simulated by convolving multi-channel impulse responses with clean speech taken from the Tohoku University and Panasonic isolated spoken word database. The modified proposed method achieved average relative error reduction rates of 22.4% in relation to the conventional CMN and 11.2% in relation to the original proposed method. By combining the modified proposed method with beamforming, the relative error reduction rates of 24.5% in relation to the conventional CMN with beamforming was achieved using only an isolated word (with duration of about 0.6 s) to estimate the spectrum of the impulse response.

So far, the spectrum of the impulse response $H(d, \omega)$ was estimated using only the corresponding utterance to be recognized with average duration of about 0.6 second. In the future, we will attempt to use multiple words to estimate the spectrum of the impulse response $H(d, \omega)$ and obtain a more accurate estimation result. In this paper, the cepstral mean was calculated using one isolated word. Our previous study [27]–[29] showed that a cepstral mean cannot be accurately estimated from a short utterance (for example, one isolated word). In the future, we will estimate the cepstral mean from multiple words and combine it with the modified proposed method. Finally, additive noise was not considered in this paper. We will attempt to evaluate our proposed methods using real-world speech data simultaneously degraded by additive noise and convolutional noise in the future.

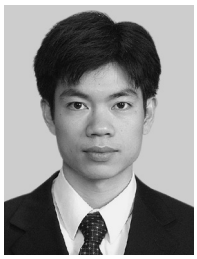
Acknowledgments

This work was partially supported by a research grant from Grant-in-Aid for Young Scientists (B) (22700169) and the MURATA Fund Grant of Hamamatsu Foundation for Science and Technology Promotion.

References

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust. Speech Signal Process., vol.29, no.2, pp.254–272, 1981.
[2] F. Liu, R. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," Proc. ARPA Speech and Nat. Language Workshop, pp.69–74, 1993.
[3] C. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation by Splitting of HMM for long reverberation," Proc. INTERSPEECH-2005, pp.277–280, 2005.
[4] C. Raut, T. Nishimoto, and S. Sagayama, "Adaptation for long convolutional distortion by maximum likelihood based state filtering approach," Proc. ICASSP-2006, vol.1, pp.1133–1136, 2006.
[5] S. Subramaniam, A.P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," IEEE Trans. Speech Audio Process., vol.4, no.5, pp.392–396, 1996.
[6] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," Proc. ICSLP-1996, pp.889–892, 1996.
[7] C. Avendano, S. Tibrewala, and H. Hermansky, "Multiresolution channel normalization for ASR in reverberation environments," Proc. EUROSPEECH-1997, pp.1107–1110, 1997.
[8] H. Hermansky, E.A. Wan, and C. Avendano, "Speech enhancement based on temporal processing," Proc. ICASSP-1995, pp.405–408, 1995.
[9] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," EURASIP Journal on Applied Signal Processing, vol.2003, no.1, pp.1074–1090, 2003.
[10] Q. Jin, Y. Pan, and T. Schultz, "Far-field speaker recognition," Proc. ICASSP-2006, vol.1, pp.937–940, 2006.
[11] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," IEEE Trans. Audio, Speech, and Language Processing, vol.15, no.7, pp.2023–2032, 2007.
[12] M. Delcroix, T. Hikichi, and M. Miyoshi, "On a blind speech dereverberation using multi-channel linear prediction," IEICE Trans. Fundamentls, vol.E89-A, no.10, pp.2837–2846, Oct. 2006.
[13] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multi-channel linear prediction," IEEE Trans. Audio, Speech, and Language Processing, vol.15, no.2, pp.430–440, Feb. 2007.
[14] I. Tashev and D. Allred, "Reverberation reduction for improved speech recognition," Proc. Hands-Free Communication and Microphone Arrays, 2005.
[15] Y. Huang and J. Benesty, "Adaptive blind channel identification: Multi-channel least mean square and Newton algorithms," ICASSP, vol.II, pp.1637–1640, 2002.
[16] Y. Huang and J. Benesty, "Adaptive multichannel least mean square and Newton algorithms for blind channel identification," Signal Process., vol.82, pp.1127–1138, Aug. 2002.
[17] Y. Huang, J. Benesty, and J. Chen, "Optimal step size of the adaptive multi-channel LMS algorithm for blind SIMO identification," IEEE Signal Process. Lett., vol.12, no.3, pp.173–175, March 2005.
[18] Y. Huang, J. Benesty, and J. Chen, Acoustic MIMO Signal Processing, Springer, 2006.
[19] M. Xu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," IEEE Trans. Signal Process., vol.43, no.12, pp.2982–2993, Dec. 1995.
[20] H. Chen, X. Cao, and J. Zhu, "Convergence of stochastic-approximation-based algorithms for blind channel identification," IEEE Trans. Inf. Theory, vol.48, no.5, pp.1214–1225, 2002.
[21] http://www.slt.atr.co.jp/ tnishi/DB/micarray/indexe.htm
[22] S. Makino, K. Niyada, Y. Mafune, and K. Kido, "Tohoku university and panasonic isolated spoken word database," J. Acoust. Soc. Jpn., vol.48, no.12, pp.899–905, Dec. 1992.
[23] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition," Proc. International Workshop on Automatic Speech Recognition and Understanding, pp.393–396, 1999.
[24] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn. (E), vol.20, no.3, pp.199–206, 1999.
[25] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE Trans. Acoust. Speech Signal Process., vol.36, no.2, pp.145–152, 1988.
[26] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," EURASIP J. APS, vol.2007, Article-ID 34013, April 2007.
[27] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN," EURASIP J. Appl. Signal Process., vol.2006, Article ID 95491, pp.1–11, 2006.
[28] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker

recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," Speech Commun., vol.49, no.6, pp.501–513, June 2007.
[29] L. Wang, S. Nakagawa, and N. Kitaoka, "Robust speech recognition by combining short-term based position-dependent CMN with conventional CMN," IEICE Trans. Inf. & Syst., vol.E91-D, no.3, pp.457–466, March 2008.



Longbiao Wang received his B.E. degree from Fuzhou University, China, in 2000 and M.E. and Dr. Eng. degrees from Toyohashi University of Technology, Japan, in 2005 and 2008, respectively. From July 2000 to August 2002, he worked at the China Construction Bank. Since 2008 he has been an assistant professor in the faculty of Engineering at Shizuoka University, Japan. His research interests include robust speech recognition, speaker recognition and source localization. He is a member of IEEE and Acoustical Society of Japan (ASJ).

Norihide Kitaoka received his B.E. and M.E. degrees from Kyoto University in 1992 and 1994, respectively, and a Dr. Engineering degree from Toyohashi University of Technology in 2000. He joined the DENSO CORPORATION, Japan in 1994. He joined the Department of Information and Computer Sciences at Toyohashi University of Technology as a Research Associate in 2001 and was a Lecturer from 2003 to 2006. Since 2006 he has been an associate professor in the Department of Media Science, Graduate School of Information Science, Nagoya University. He was a visiting associate professor of Nanyang Technological University, Singapore, in 2009. His research interests include speech processing, speech recognition, and spoken dialog. He is a member of ISCA, ASJ, Information Processing Society of Japan (IPSJ), and Japan Society for Artificial Intelligence (JSAI).

Seiichi Nakagawa received a Dr. of Eng. degree from Kyoto University in 1977. He joined the faculty of Kyoto University, in 1976, as a Research Associate in the Department of Information Sciences. He moved to Toyohashi University of Technology in 1980. From 1980 to 1983 he was an Assistant Professor, and from 1983 to 1990 he was an Associate Professor. Since 1990 he has been a Professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi. From 1985 to 1986, he was a Visiting Scientist in the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, USA. He received the 1997/2001 Paper Award from the IEICE and the 1988 JC Bose Memorial Award from the Institution of Electro. Telecomm. Engrs. His major interests in research include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence. He is a fellow of IPSJ.