Chapter 7

Dereverberation Based on Spectral Subtraction by Multi-Channel LMS Algorithm for Hands-Free Speech Recognition

Longbiao Wang, Kyohei Odani, Atsuhiko Kai, Norihide Kitaoka and Seiichi Nakagawa

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48430

1. Introduction

In a distant-talking environment, channel distortion drastically degrades speech recognition performance because of a mismatch between the training and testing environments. The current approach focusing on automatic speech recognition (ASR) robustness to reverberation and noise can be classified as speech signal processing [1, 4, 5, 14], robust feature extraction [10, 20], and model adaptation [3, 25].

In this chapter, we focus on speech signal processing in the distant-talking environment. Because both the speech signal and the reverberation are nonstationary signals, dereverberation to obtain clean speech from the convolution of nonstationary speech signals and impulse responses is very hard work. Several studies have focused on mitigating the above problem [8, 9, 11, 12]. [1] explored a speech dereverberation technique whose principle was the recovery of the envelope modulations of the original (anechoic) speech. They applied a technique that they originally developed to treat background noise [11] to the dereverberation problem. [7] proposed a novel approach for multimicrophone speech dereverberation. 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 by [16, 17]. 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. [18] proposed a novel dereverberation method utilizing multi-step forward linear prediction. They estimated the linear prediction coefficients in a time domain and suppressed the amplitude of late reflections through spectral subtraction in a spectral domain.

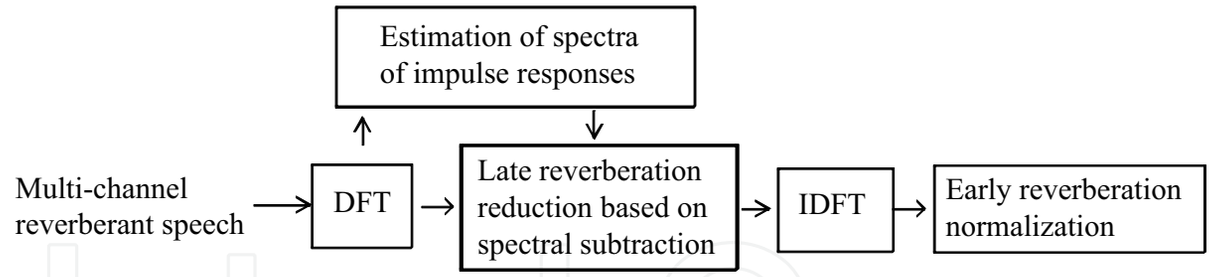


Figure 1. Schematic diagram of blind dereverberation method.

In this chapter, we propose a robust distant-talking speech recognition method based on spectral subtraction (SS) employing the multi-channel least mean square (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 information from the dereverberation operation, the dereverberation reduction in a power spectral domain provides robustness against certain errors that the conventional sensitive inverse filtering method cannot achieve [18]. The compensation parameter (that is, the spectrum of the impulse response) for spectral subtraction is required. An adaptive MCLMS algorithm was proposed to blindly identify the channel impulse response in a time domain [12–14]. In this chapter, we extend the method to blindly estimate the spectrum of the impulse response for spectral subtraction in a frequency domain. The early reverberation is normalized by CMN [6]. Power SS is the most commonly used SS method. A previous study has shown that generalized SS (GSS) with a lower exponent parameter is more effective than power SS for noise reduction [26]. In this chapter, both of power SS and GSS are employed to suppress late reverberation. A diagram of the proposed method is shown in Fig. 1.

In this chapter, we also investigate the robustness of the power SS-based dereverberation under various reverberant conditions for large vocabulary continuous speech recognition (LVCSR). We analyze the effect factors (numbers of reverberation windows and channels, length of utterance, and the distance between sound source and microphone) of compensation parameter estimation for dereverberation based on power SS in a simulated reverberant environment.

The remainder of this paper is organized as follows. Section 2 describes the outline of blind dereverberation 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 described in Section 3. Section 4 describes the experimental results of hands-free speech recognition in both simulated and real reverberant environments. Finally, Section 5 summarizes the paper.

2. Outline of blind dereverberation

2.1. Dereverberation based on power SS

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

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

where * denotes the convolution operation. In this chapter, 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 [16, 17]

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

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 [25].

[17] proposed a far-field speaker recognition based on spectral subtraction. In this method, 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 $|1 - 0.9e^{j\omega}|$ [17]. However, the drawback of this approach is that the optimum parameters α, β for the spectral subtraction are 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 chapter, 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 Section 3. Assuming that phases of different frames is noncorrelated for simplification, the power spectrum of Eq. (4) can be approximated as

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

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 chapter, the late reverberation is reduced based on the power SS, while the early reverberation is normalized by CMN at the feature extraction stage. A diagram of the proposed method is shown in Fig. 1. SS is used to prevent the estimated power spectrum obtained by reducing the late reverberation from being a negative value; the estimated power spectrum $|\hat{X}(f,\omega)|^2$ obtained by reducing the late reverberation then becomes

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

where $|\hat{X}(f,\omega)|^2 = |\hat{S}(f,\omega)|^2|\hat{H}(0,\omega)|^2$, $|\hat{S}(f,\omega)|^2$ is the spectrum of estimated clean speech, $\hat{H}(f,\omega)$ is the estimated STFT of the impulse response. To estimate the power spectra of the impulse responses, we extended the Multi-channel LMS algorithm for identifying the impulse responses in a time domain [14] to a frequency domain in Section 3.2.

2.2. Dereverberation based on GSS

Previous studies have shown that GSS with an arbitrary exponent parameter is more effective than power SS for noise reduction [26]. In this chapter, we extend GSS to suppress late reverberation. Instead of the power SS-based dereverberation given in Eq. (7), GSS-based dereverberation is modified as

$$|\hat{X}(f,\omega)|^{2n} \approx max\{|X(f,\omega)|^{2n} - \alpha \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{X}(f-d,\omega)|^{2n}|\hat{H}(d,\omega)|^{2n}\}}{|\hat{H}(0,\omega)|^{2n}}, \beta \cdot |X(f,\omega)|^{2n}\}, \quad (8)$$

where n is the exponent parameter. For power SS, the exponent parameter n is equal to 1. In this chapter, the exponent parameter n is set to 0.1 as this value yielded the best results [26].

The methods given in Eq. (7) and Eq. (8) are referred to *power SS-based* and *GSS-based dereverberation methods*, respectively.

2.3. Dereverberation and denoising based on GSS

The precision of impulse response estimation is drastically degraded when the additive noise is present. We present a dereverberation and denoising based on GSS. A diagram of the

processing method is shown in Fig. 2. At first, the spectrum of additive noise is estimated and noise reduction is performed. Then the reverberation is suppressed using the estimated spectra of impulse responses. When additive noise is present, the power spectrum of Eq. (6) becomes

$$|X(f,\omega)|^2 \approx |S(f,\omega)|^2|H(0,\omega)|^2 + \sum_{d=1}^{D-1} |S(f-d,\omega)|^2|H(d,\omega)|^2 + |N(f,\omega)|^2, \tag{9}$$

where $N(f,\omega)$ is the spectrum of noise $n(t)$. To suppress the noise and reverberation

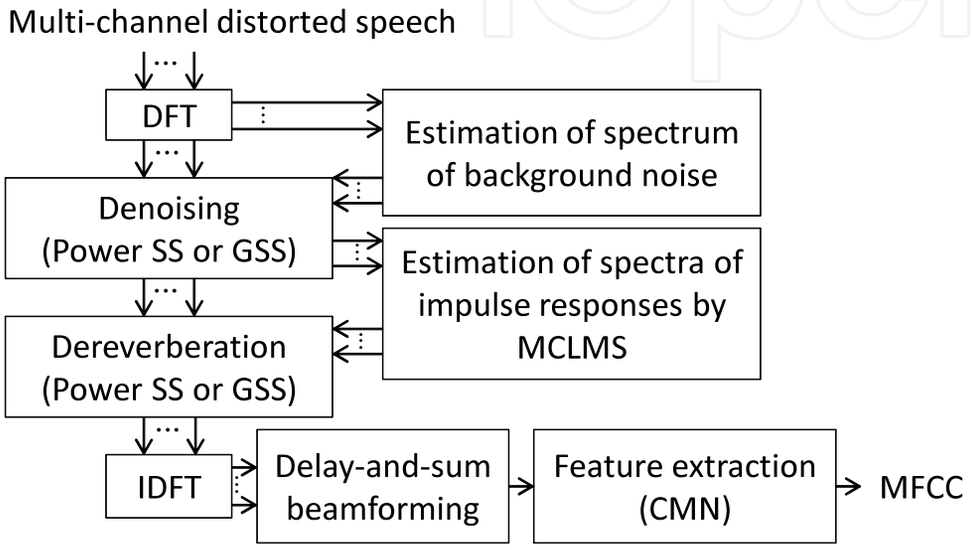


Figure 2. Schematic diagram of an SS-based dereverberation and denoising method.

simultaneously, Eq. (8) is modified as

$$|\hat{X}(f,\omega)|^{2n} \approx max\{|\hat{X}_N(f,\omega)|^{2n} - \alpha_1 \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{X}(f-d,\omega)|^{2n}|\hat{H}(d,\omega)|^{2n}\}}{|\hat{H}(0,\omega)|^{2n}}, \beta_1 \cdot |\hat{X}_N(f,\omega)|^{2n}\}, \tag{10}$$

$$|\hat{X}_N(f,\omega)|^{2n} \approx max\{|X(f,\omega)|^{2n} - \alpha_2 \cdot |\hat{N}(\omega)|^{2n}, \beta_2 \cdot |X(f,\omega)|^{2n}\}, \tag{11}$$

where $\hat{N}(\omega)$ is the mean of noise spectrum $N(f,\omega)$, and $\hat{X}_N(f,\omega)$ is the spectrum obtained by subtracting the spectrum of the observed speech from the estimated mean spectrum of noise $\hat{N}(\omega)$ [1]. In this paper, we set parameter β_1 equal to β_2.

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

An adaptive multi-channel LMS algorithm for blind Single-Input Multiple-Output (SIMO) system identification in time domain was proposed by [13, 14].

[1] In this study, stationary noise is assumed.

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. 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 [2] , 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

$$x_i * h_j = s * h_i * h_j = x_j * h_i, \; i, j = 1, 2, \cdots, N, i \neq j, \tag{12}$$

and have the following relation at time t:

$$\mathbf{x}_i^T(t)\mathbf{h}_j(t) = \mathbf{x}_j^T(t)\mathbf{h}_i(t), \quad i, j = 1, 2, \cdots, N, i \neq j, \tag{13}$$

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

$$\begin{aligned} \mathbf{x}_n(t) = [x_n(t) \;\; x_n(t-1) \;\; \cdots \;\; x_n(t-L+1)]^T, \\ n = 1, 2, \cdots, N, \end{aligned} \tag{14}$$

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. (13) by $\mathbf{x}_n(t)$ and taking expectation yields,

$$\begin{aligned} &\mathbf{R}_{x_i x_i}(t+1)\mathbf{h}_j(t) = \mathbf{R}_{x_i x_j}(t+1)\mathbf{h}_i(t), \\ &i, j = 1, 2, \cdots, N, i \neq j, \end{aligned} \tag{15}$$

where $\mathbf{R}_{x_i x_j}(t+1) = E\{\mathbf{x}_i(t+1)\mathbf{x}_j^T(t+1)\}$. Eq. (15) 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

$$\begin{aligned} &\sum_{i=1, i\neq j}^{N} \mathbf{R}_{x_i x_i}(t+1)\mathbf{h}_j(t) = \sum_{i=1, i\neq j}^{N} \mathbf{R}_{x_i x_j}(t+1)\mathbf{h}_i(t), \\ &j = 1, 2, \cdots, N. \end{aligned} \tag{16}$$

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

$$\mathbf{R}_{x+}(t+1)\mathbf{h}(t) = 0, \tag{17}$$

where

[2] 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.

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

$$\tilde{\mathbf{R}}_{x+}(t+1) = \begin{bmatrix} \sum_{n\neq 1}\tilde{\mathbf{R}}_{x_n x_n}(t+1) & -\tilde{\mathbf{R}}_{x_2 x_1}(t+1) & \cdots & -\tilde{\mathbf{R}}_{x_N x_1}(t+1) \\ -\tilde{\mathbf{R}}_{x_1 x_2}(t+1) & \sum_{n\neq 2}\tilde{\mathbf{R}}_{x_n x_n}(t+1) & \cdots & -\tilde{\mathbf{R}}_{x_N x_2}(t+1) \\ \vdots & \vdots & \ddots & \vdots \\ -\tilde{\mathbf{R}}_{x_1 x_N}(t+1) & -\tilde{\mathbf{R}}_{x_2 x_N}(t+1) & \cdots & \sum_{n\neq N}\tilde{\mathbf{R}}_{x_n x_n}(t+1) \end{bmatrix}, \tag{22}$$

$$\mathbf{h}(t) = [\mathbf{h}_1(t)^T \quad \mathbf{h}_2(t)^T \quad \cdots \quad \mathbf{h}_N(t)^T]^T, \tag{19}$$

$$\mathbf{h}_n(t) = [h_n(t,0) \quad h_n(t,1) \quad \cdots \quad h_n(t,L-1)]^T, \tag{20}$$

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:

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

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 $t+1$

$$J(t+1) = \|\mathbf{e}(t+1)\|^2 = \mathbf{e}(t+1)^T\mathbf{e}(t+1). \tag{23}$$

By minimizing the cost function J of Eq. (23), 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 [12, 14]. Among these methods, the VSS-UMCLMS achieves a nice balance between complexity and convergence speed [14]. 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 chapter, 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

$$\Delta J(t+1) \approx \frac{2\tilde{\mathbf{R}}_{x+}(t+1)\hat{\mathbf{h}}(t)}{\|\hat{\mathbf{h}}(t)\|^2} \tag{24}$$

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

$$\hat{\mathbf{h}}(t+1) = \hat{\mathbf{h}}(t) - 2\mu\tilde{\mathbf{R}}_{x+}(t+1)\hat{\mathbf{h}}(t), \tag{25}$$

which is theoretically equivalent to the adaptive algorithm proposed by [2] although the cost functions are defined in different ways in these two adaptive blind SIMO identification algothrithms. In Eq. (25), 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 $\mathbf{h}$ [2].

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

$$\mu_{opt}(t+1) = \frac{\hat{\mathbf{h}}^T(t)\Delta J(t+1)}{\|\Delta J(t+1)\|^2}. \tag{26}$$

The details of the VSS-UMCLMS were described in [14].

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 Section 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:

$$\mathbf{X}_i^T(f)\mathbf{H}_j(f) = \mathbf{X}_j^T(f)\mathbf{H}_i(f), \quad i,j = 1,2,...,N, \quad i \neq j, \tag{27}$$

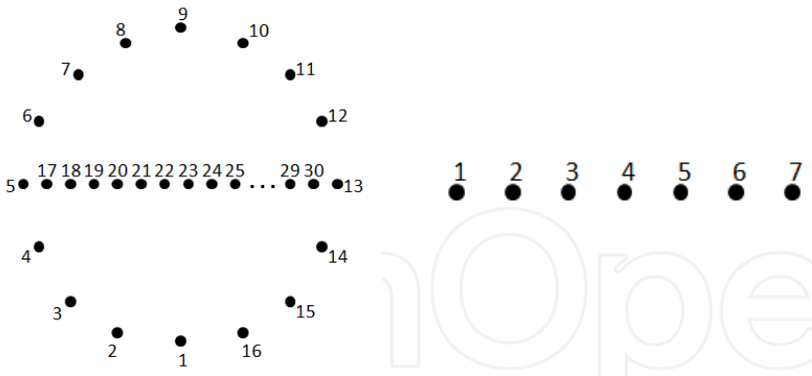
Where $\mathbf{X}_n(f) = [X_n(f) \quad X_n(f-1) \quad ... \quad X_n(f-D+1)]^T$ is a D-dimention vector of spectra of the distorted speech received from the *n-th* channel at frame f, $X_n(f)$ is the spectrum of the distorted speech received from the *n-th* channel at frame f for frequency ω, $\mathbf{H}_n(f) = [H_n(f,0) \quad H_n(f,1) \quad ... \quad H_n(f,d) \quad ... \quad H_n(f,D-1)]^T, d = 0,1,...,D-1$ 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. (27) in place of Eq. (13), the spectra of the impulse responses can be blindly estimated by the VSS-UMCLMS mentioned in Section 3.1.2.

4. Experiments

4.1. Experimental setup

The proposed dereverberation method based on spectral subtraction is evaluated on an isolated word recognition task in a simulated reverberant environment, and a large vocabulary continuous speech recognition task in both a simulated reverberant environment and a real reverberant environment, respectively.



(a) RWCP (b) CENSREC-4

Figure 3. Illustration of microphone array.

(a) RWCP database

Array no	Array type	Room	Angle	RT60 (S)
1	linear	Echo room (panel)	150°	0.30
2	circle	Echo room (cylinder)	30°	0.38
3	linear	Tatami-floored room (S)	120°	0.47
4	circle	Tatami-floored room (S)	120°	0.47
5	circle	Tatami-floored room (L)	90°	0.60
6	circle	Tatami-floored room (L)	130°	0.60
7	linear	Conference room	50°	0.78
8	linear	Echo room (panel)	70°	1.30

(b) CENSREC-4 database

Array no	Room	Room size	RT60 (s)
9	Office	9.0 Ą 6.0 m	0.25
10	Japanese style room	3.5 Ą 2.5 m	0.40
11	Lounge	11.5 Ą 27.0 m	0.50
12	Japanese style bath	1.5 Ą 1.0 m	0.60
13	Living room	7.0 Ą 3.0 m	0.65
14	Meeting room	7.0 Ą 8.5 m	0.65
15	Elevator hall	11.5 Ą 6.5 m	0.75

RT60 (second): reverberation time in room; S: small; L: large

Table 1. Details of recording conditions for impulse response measurement

4.1.1. Experimental setup for isolated word recognition task

Multi-channel distorted speech signals simulated by convolving multi-channel impulse responses with clean speech were used to create artificial reverberant speech. Six kinds of multi-channel impulse responses measured in various acoustical reverberant environments were selected from the Real World Computing Partnership (RWCP) sound scene database [23]. Table 1 lists the details of recording conditions (impulse responses with array no 3-8 in RWCP

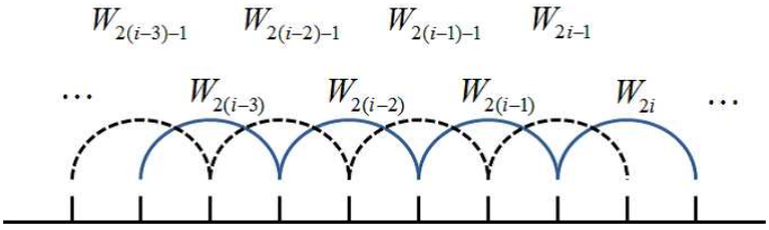


Figure 4. Illustration of the analysis window for spectral subtraction.

database were used in the isolated word recognition task). The illustration of microphone array is shown in Fig. 3. A four-channel circle or linear microphone array was taken from a circle + 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*.

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 [21]. 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 [22]) were trained using 27,992 utterances read by 175 male speakers from the Japanese Newspaper Article Sentences (JNAS) corpus [15]). 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 was 256 (21.3 *ms*), and the rate of overlap was 1/2. An illustration of the analysis window is shown in Fig. 4. 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 [3]. The spectrum of the impulse response $\hat{H}(d,\omega)$ 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

[3] Eq. (27) is true when using a skip window and the spectrum of the impulse response can be blindly estimated.

microphone	SONY ECM-C10
A/D board	Tokyo Electron device TD-BD-16ADUSB
recording room size [m]	7.1(D) Ą 3.3(W) Ą 2.5(H)
number of speakers	5 male speakers
number of utterances	100 utterances (about 20 utterances per speaker)
background noise	electric fan
sampling frequency	16 kHz
quantization bit rate	16 bits

Table 2. Conditions for recording in real environment.

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%.

4.1.2. Experimental setup for LVCSR task

In this study, both the artificial reverberant speech and real reverberant speech were used to evaluate our proposed method.

For artificial reverberant speech, multi-channel distorted speech signals simulated by convolving multi-channel impulse responses with clean speech were used. Fifteen kinds of multi-channel impulse responses measured in various acoustical reverberant environments were selected from the real world computing partnership (RWCP) sound scene database [23] and the CENSREC-4 database [24]. Table 1 lists the details of 15 recording conditions. The illustration of microphone array is shown in Fig. 3. For RWCP database, a 2–8 channel circle or linear microphone array was taken from a circle + linear microphone array (30 channels). The circle type microphone array had a diameter of 30 cm. The microphones of the linear microphone array were located at 2.83 cm intervals. Impulse responses were measured at several positions 2 m from the microphone array. For the CENSREC-4 database, 2 or 4 channel microphones were taken from a linear microphone array (7 channels) with the two microphones located at 2.125 cm intervals. Impulse responses were measured at several positions 0.5 m from the microphone array. The Japanese Newspaper Article Sentences (JNAS) corpus [15] was used as clean speech. Hundred utterances from the JNAS database convolved with the multi-channel impulse responses shown in Table 1 were used as test data. The average time for all utterances was about 5.8 s.

For reverberant speech in a real environment, we recorded multi-channel speech degraded simultaneously by background noise and reverberation. Table 2 gives the conditions and content of the recordings. One hundred utterances from the JNAS corpus, uttered by five male speakers seated on the chairs labeled A to E in Fig. 5, were recorded by a multi-channel recording device. The heights of the microphone array and the utterance position of each speaker were about 0.8 m and 1.0 m, respectively. An electric fan with high air volume located behind the speaker in position A was used as background noise. An average SNR of the speech was about 18 dB. We used amicrophone array with 9 channels (Fig. 5) and a pin microphone to record speech in the distant-talking environment and close-talking environment, respectively.

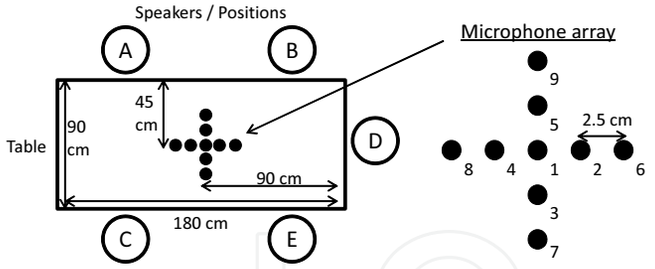


Figure 5. Illustration of recording settings and microphone array in real environment

Sampling frequency	16 kHz
Frame length	25 ms
Frame shift	10 ms
Acoustic model	5 states, 3 output probability left-to-right triphone HMMs
Feature space	25 dimensions with CMN (12 MFCCs + Δ + Δpower)

Table 3. Conditions for large vocabulary continuous speech recognition

method	Power SS		GSS	
	DN	DR	DN	DR
analysis window	Hamming			
window length	32 ms			
window shift	16 ms			
noise overestimation factor α	α_2= 3.0	α_1= 1.0	α_1=α_2= 0.1	
spectral floor parameter β	β_1=β_2=0.15			

Table 4. Conditions for SS-based denoising and dereverberation. “DN”: denoising. “DR”: dereverberation.

Table 3 gives the conditions for speech recognition. The acoustic models were trained with the ASJ speech databases of phonetically balanced sentences (ASJ-PB) and the JNAS. In total, around 20K sentences (clean speech) uttered by 132 speakers were used for each gender. Table 4 gives the conditions for SS-based denoising and dereverberation. The parameters shown in Table 4 were determined empirically. For SS-based dereverberation method without background noise, the parameter α was equal to α_1 and β was equal to β_1. The number of reverberant windows D was set to 6 (192 ms). An illustration of the analysis window is shown in Fig. 4. An open-source LVCSR decoder software "Julius" [19] that is based on word trigram and triphone context-dependent HMMs is used.

4.2. Experimental results

4.2.1. Isolated word recognition results

Table 5 shows the isolated word recognition results in a simulated reverberant environment. "Distorted speech #" in Table 5 corresponds to "array no" in Table 1. Delay-and-sum

Distorted speech #	CMN	Power SS-based dereverberation
3	69.4	76.0
4	73.2	80.6
5	71.4	80.3
6	71.8	78.6
7	67.7	74.4
8	63.1	71.2
Ave.	69.4	76.9

Delay-and-sum beamforming was performed for all methods

Table 5. Isolated word recognition results (%).

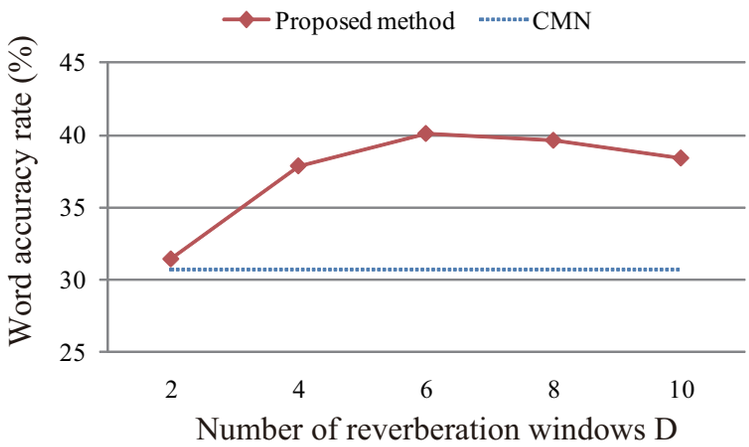


Figure 6. Effect of the number of reverberation windows D on power SS-based dereverberation for speech recognition.

beamforming [27] is performed for all methods in this chapter. The conventional CMN combined with delay-and-sum beamforming was used as a baseline.

The power SS-based dereverberation method by Eq. (7) improved speech recognition significantly compared with CMN for all severe reverberant conditions. The reason was that the proposed method compensated for both the late and early reverberation. The proposed method achieved an average relative error reduction rate of 24.5% in relation to conventional CMN with beamforming.

4.2.2. LVCSR results

(a) Effect factor analysis of power SS-based dereverberation in the simulated reverberant environment

In this section, we describe the use of four microphones to estimate the spectrum of the impulse responses without a particular explanation. Delay-and-sum beamforming (BF) was performed on the 4-channel dereverberant speech signals. For the proposed method, each speech channel was compensated by the corresponding estimated impulse response. Preliminary experimental results for isolated word recognition showed that the power SS-based dereverberation method significantly improved the speech recognition performance significantly compared with traditional CMN with beamforming. In this section, we

Array no #	Number of reverberation windows D				
	2	4	6	8	10
1	**81.45**	80.43	79.94	79.67	79.98
2	43.89	55.71	**57.69**	54.06	51.98
3	23.40	32.02	**33.46**	33.29	32.81
4	28.77	38.42	39.69	**39.88**	38.92
5	22.89	30.26	33.34	**33.59**	31.71
6	21.01	27.46	**31.79**	31.32	28.97
7	15.89	20.55	23.32	**23.92**	22.54
8	14.26	17.94	**21.41**	21.12	20.24
Ave	31.44	37.85	40.08	39.61	38.39

The results with bold font indicate the best result corresponding to each array

Table 6. Detail results based on different number of reverberation windows D and reverberant environments (%)

	Linear array	Circle array
2 channels	17, 29	1, 9
4 channels	17, 21, 25, 29	1, 5, 9, 13
8 channels	17, 19, 21, 23, 25, 27, 29, 30	1, 3, 5, 7, 9, 11, 13, 15, 17

Table 7. Channel number corresponding to Fig. 3(a) using for dereverberation and denoising (RWCP database)

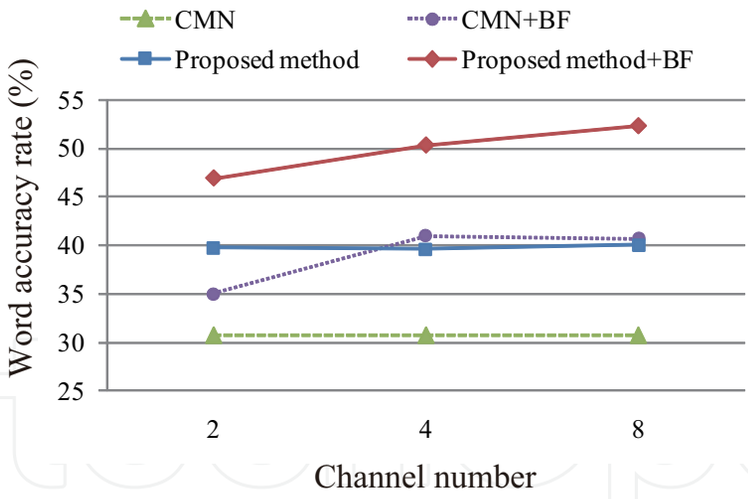


Figure 7. Effect of the number of channels on power SS-based dereverberation for speech recognition.

evaluated the power SS-based dereverberation method for LVCSR and analyzed the effect factor (number of reverberation windows D in Eq. (7), channel number, and length of utterance) for compensation parameter estimation based on power SS using RWCP database. The word accuracy rate for LVCSR with clean speech was 92.6%.

The effect of the number of reverberation windows on speech recognition is shown in Fig. 6. The detail results based on different number of reverberation windows D and reverberant environments (that is, different reverberation times) were shown in Table 6. The results shown on Fig. 6 and Table 6 were not performed delay-and-sum beamforming. The results show

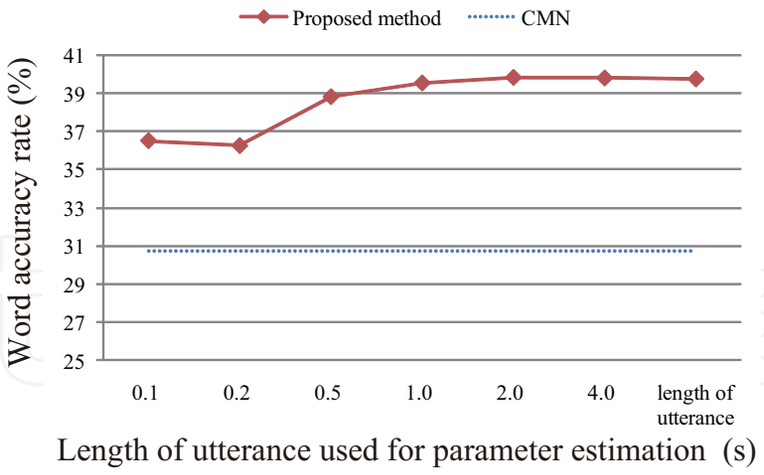


Figure 8. Effect of length of utterance used for parameter estimation on power SS-based dereverberation for speech recognition.

that the optimal number of reverberation windows D depends on the reverberation time. The best average result of all reverberant speech was obtained when D equals 6. The speech recognition performance with the number of reverberation windows between 4 and 10 did not vary greatly and was significantly better than the baseline.

We analyzed the influence of the number of channels on parameter estimation and delay-and-sum beamforming. Besides four channels, two and eight channels were also used to estimate the compensation parameter and perform beamforming. Channel numbers corresponding to Fig. 3(a) shown in Table 7 were used. The results are shown in Fig. 7. The speech recognition performance of the SS-based dereverberation method without beamforming was hardly affected by the number of channels. That is, the compensation parameter estimation is robust to the number of channels. Combined with beamforming, the more channels that are used and the better is the speech recognition performance.

Thus far, the whole utterance has been used to estimate the compensation parameter. The effect of the length of utterance used for parameter estimation was investigated, with the results shown in Fig. 8. The longer the length of utterance used, the better is the speech recognition performance. Deterioration in speech recognition was not experienced with the length of the utterance used for parameter estimation greater than 1 s. The speech recognition performance of the SS-based dereverberation method is better than the baseline even if only 0.1 s of utterance is used to estimate the compensation parameter.

We also compared the power SS-based dereverberation method on LVCSR in different simulated reverberant environments. The experimental results shown in Fig. 9. Naturally, the speech recognition rate deteriorated as the reverberation time increased. Using the SS-based dereverberation method, the reduction in the speech recognition rate was smaller than in conventional CMN, especially for impulse responses with a long reverberation time. For RWCP database, the SS-based dereverberation method achieved a relative word recognition error reduction rate of 19.2% relative to CMN with delay-and-sum beamforming. We also conducted an LVCSR experiment with SS-based dereverberation under different reverberant conditions (CENSREC-4), with the reverberation time between 0.25 and 0.75 s and the distance between microphone and sound source 0.5 m. A similar trend to the above results was observed. Therefore, the SS-based dereverberation method is robust to various reverberant

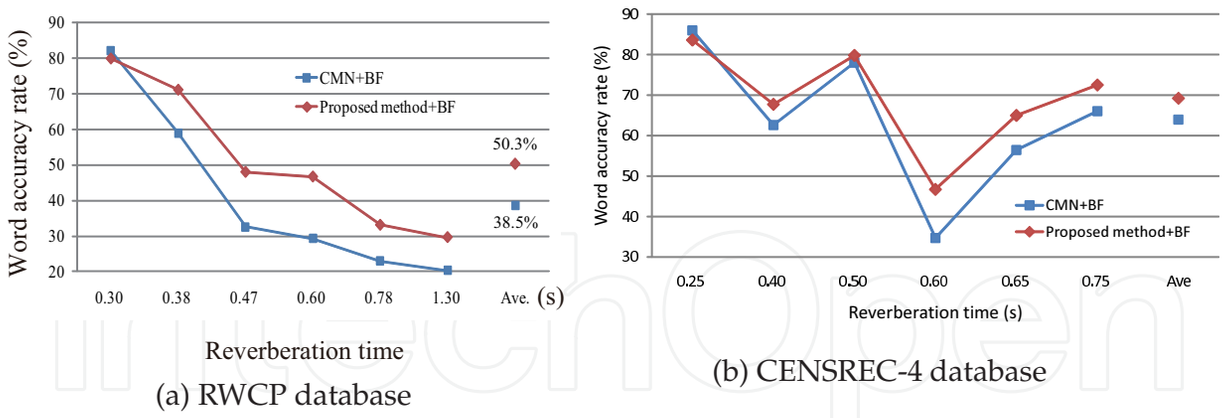


Figure 9. Word accuracy for LVCSR in different simulated reverberant environments.

conditions for both isolated word recognition and LVCSR. The reason is that the SS-based dereverberation method can compensate for late reverberation through SS using an estimated power spectrum of the impulse response.

(b) Results of GSS-based method in the simulated reverberant environment

In this section, reverberation and noise suppression using only 2 speech channels is described. In both power SS-based and GSS-based dereverberation methods, speech signals from two microphones were used to estimate blindly the compensation parameters for the power SS and GSS (that is, the spectra of the channel impulse responses), and then reverberation was suppressed by SS and the spectrum of dereverberant speech was inverted into a time domain. Finally, delay-and-sum beamforming was performed on the two-channel dereverberant speech.

The results of power SS-based method and the GSS-based method without background noise were compared in Table 8. "Distorted speech #" in Table 8 corresponds to "array no" in Table 1. The speech recognition performance was drastically degraded under reverberant conditions because the conventional CMN did not suppress the late reverberation. Delay-and-sum beamforming with CMN (41.91%) could not markedly improve the speech recognition performance because of the small number of microphones and the small distance between the microphone pair. In contrast, the power SS-based dereverberation using Eq. (7) markedly improved the speech recognition performance. The GSS-based dereverberation using Eq. (8) improved speech recognition performance significantly compared with the power SS-based dereverberation and CMN for all reverberant conditions. The GSS-based method achieved an average relative word error reduction rate of 31.4% compared to the conventional CMN and 9.8% compared to the power SS-based method.

Table 9 shows the speech recognition results for the power SS and GSS-based denoising and dereverberation methods for the simulated noisy and reverberant speech. "Distorted speech #", "DN" and "DNR" in Table 9 denote the "array #" in Table 1, "denoising", and "denoising and dereverberation", respectively. The speech recognition performance of conventional CMN was drastically degraded owing to the noisy and reverberant conditions and the fact that CMN did not suppress the late reverberation. The power SS-based DN improved speech recognition performance significantly compared to the CMN for all reverberant conditions. The GSS-based DN using Eq. (11), however, did not improve the speech recognition performance compared to the power SS-based DN. On the other hand, the power SS-based DNR achieved a marked improvement in the speech recognition performance compared with

Distorted speech #	CMN only	Power SS-based method	GSS-based method
2	44.35	63.34	65.95
4	27.59	40.79	49.16
5	25.61	42.55	49.29
11	73.90	79.26	80.77
12	27.06	42.28	45.38
13	29.62	50.78	56.13
15	65.24	71.67	74.35
Ave.	41.91	55.81	60.15

Delay-and-sum beamforming was performed for all methods

Table 8. Comparison of Word accuracy for LVCSR with power SS-based method and GSS-based method in the simulated reverberant environment (%)

Distorted	CMN	Power SS		GSS	
Speech #	only	DN	DNR	DN	DNR
1	28.2	37.4	48.8	30.3	48.3
2	16.0	25.9	33.5	18.8	36.3
3	9.5	21.3	31.3	13.9	32.8
4	55.8	72.2	69.9	60.4	68.2
5	17.2	24.4	32.0	20.9	37.7
6	26.1	32.8	45.3	30.0	51.7
7	54.4	64.6	66.5	57.7	68.8
Average	29.6	39.8	46.7	33.1	49.1

Delay-and-sum beamforming was performed for all methods

Table 9. Word accuracy for LVCSR with the simulated noisy reverberant speech (%).

that of CMN. The GSS-based DNR using Eq. (10) improved speech recognition performance significantly compared to both the CMN method and the power SS-based DNR for almost all reverberant conditions.

(c) Results in the real noisy reverberant environment

Table 10 shows the speech recognition results for the real noisy reverberant speech under the same conditions as the simulated noisy reververant speech. The word accuracy rate for close-talking speech recorded in a real environment was 88.3%. We investigated the best channel combination in the real environment and the best speech recognition performance was obtained when channels 6, 7, 8, and 9 described in Fig. 5 were used. Therefore, this channel combination was used in this study. Power SS-based DN and GSS-based DN achieved a smaller improvement in recognition performance compared with the simulated noisy reverberant environment because the type of background noise in the real environment was different from that in the simulated environment. On the other hand, the power SS-based DNR markedly improved the speech recognition performance compared to CMN. The GSS-based DNR improved speech recognition performance significantly compared to

Speakers / Position	CMN only	Power SS		GSS	
		DN	DNR	DN	DNR
A	60.2	67.7	78.9	64.7	79.5
B	75.6	72.2	78.5	72.5	83.2
C	67.4	63.2	69.4	66.7	77.5
D	59.1	53.9	74.9	60.8	78.7
E	42.9	51.0	62.8	50.0	61.7
Average	60.9	61.6	73.1	62.9	76.2

Delay-and-sum beamforming was performed for all methods

Table 10. Word accuracy for LVCSR with the real noisy reverberant speech (%).

both the CMN method and the power SS-based DNR for almost all speakers. The GSS-based DNR achieved an average relative word error reduction rate of 39.1% and 11.5% compared to conventional CMN and power SS-based DNR, respectively. These results show that our proposed method is also effective in a real environment under the same denoising and dereverberation conditions as the simulated noisy reverberant environment.

5. Conclusion

In this chapter, we proposed a blind spectral subtraction based dereverberation method for hands-free speech recognition method. We treated the late reverberation as additive noise, and a noise reduction technique based on spectral subtraction was applied to compensate for the late reverberation. The early reverberation was normalized by CMN. The time-domain MCLMS algorithm was extended to blindly estimate the spectrum of the impulse response for spectral subtraction in a frequency domain. We evaluated our proposed methods on isolated word recognition task and LVCSR task. The proposed spectral subtraction based on multi-channel LMS significantly outperformed than the conventional CMN. For isolated word recognition task, a relative error reduction rate of 24.5% in relation to the conventional CMN was achieved. For LVCSR task without background noise, the proposed method achieved an average relative word error reduction rate of 31.5% compared to conventional CMN in the simulated reverberant environment. We also presented a denoising and dereverberation method based on spectral subtraction and evaluated it in both the simulated noisy reverberant environment and the real noisy reverberant environment. The GSS-based method achieved an average relative word error reduction rate of 39.1% and 11.5% compared to conventional CMN and power SS-based method, respectively. These results show that our proposed method is also effective in a real noisy reverberant environment.

In this chapter, we also investigated the effect factors (numbers of reverberation windows and channels, and length of utterance) for compensation parameter estimation. We reached the following conclusions: 1) the speech recognition performance with the number of reverberation windows between 4 and 10 did not vary greatly and was significantly better than the baseline; 2) the compensation parameter estimation was robust to the number of channels; and 3) degradation of speech recognition did not occur with the length of utterance used for parameter estimation longer than 1 *s*. We also compared the SS-based dereverberation method on LVCSR in different simulated reverberant environments. A similar trend was observed.

Author details

Longbiao Wang, Kyohei Odani and Atsuhiko Kai
Shizuoka University, Japan

Norihide Kitaoka
Nagoya University, Japan

Seiichi Nakagawa
Toyohashi University of Technology, Japan

6. References

[1] Avendano, C. & Hermansky, H. (1996). Study on the dereverberation of speech based on temporal envelope filtering. *Proceedings of ICSLP-1996*, pp. 889-892, Philadelphia, USA, October 1996.

[2] Chen, H., Cao, X., & Zhu, J. (2002). Convergence of stochastic-approximation-based algorithms for blind channel identification. *IEEE Trans. Information Theory*, Vol. 48, 2002, pp. 1214-1225.

[3] Couvreur, L. & Couvreur, C. (2004). Blind model selection for automatic speech recognition in reverberant environments. *Journal of VLSI Signal Processing*, Vol. 36, No. 2-3, February/March 2004, pp. 189-203.

[4] Delcroix, M., Hikichi, T. & Miyoshi, M. (1994). On a blind speech dereverberation algorithm using multi-channel linear prediction. *IEEE Transations on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. E89-A, No. 10, October 2006, pp. 2837-2846.

[5] Delcroix, M., Hikichi, T. & Miyoshi, M. (1994). Precise dereverberation using multi-channel linear prediction. *IEEE Transations on Audio, Speech, and Language Processing*, Vol. 15, No. 2, February 2007, pp. 430-440.

[6] Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acous. Speech Signal Processing*, Vol. 29, No. 2, 1981, pp. 254-272.

[7] Gannot, S. & Moonen, M. (2003). Subspace methods for multimicrophone speech dereverberation. *EURASIP Journal on Applied Signal Processing*, October 2003, pp. 1074-1090.

[8] Gillespie, B. W., Malvar, H. S. & Florencio, D. A. F. (2001). Speech dereverberation via maximum-kurtosis subband adaptive filtering, *Proceedings of ICASSP-2001*, Vol. 6, pp. 3701-3704, Salt Lake City, USA, May 2001.

[9] Habets, E. A. P. (2004). Single-channel speech dereverberation based on spectral subtraction, *Proceedings of the 15th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC-2004)*, pp. 250-254, Veldhoven, Netherlands, November 2004.

[10] Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. *IEEE Transations on Speech and Audio Processing*, Vol. 2, No. 4, October 1994, pp. 578-589.

[11] Hermansky, H., Wan, E. A. & Avendano, C. (1995). Speech enhancement based on temporal processing, *Proceedings of ICASSP-1995*, pp. 405-408, Detroit, USA, May 1995.

[12] Huang, Y. & Benesty, J. (2002). Adaptive multichannel least mean square and Newton algorithms for blind channel identification. *Signal Processing*, Vol. 82, No. 8, August 2002, pp. 1127-1138.

[13] Huang, Y., Benesty, J. & Chen, J. (2005). Optimal step size of the adaptive multi-channel LMS algorithm for blind SIMO identification. *IEEE Signal Processing Letters*, Vol. 12, No. 3, March 2005, pp. 173-176.

[14] Huang, Y., Benesty, J. & Chen, J. (2006). *Acoustic MIMO Signal Processing*, Springer-Verlag, ISBN 978-3-540-37630-9, Berlin, Germany.
[15] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. & Itahashi, S. (1999). JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of the Acoustical Society of Japan (E)*, Vol. 20, No. 3, May 1999, pp. 199-206.
[16] Jin, Q., Pan, Y. & Schultz, t. (2006). Far-field speaker recognition, *Proceedings of ICASSP-2006*, pp. 937-940, Toulouse, France, May 2006.
[17] Jin, Q., Schultz, t. & Waibel, A. (2007). Far-field speaker recognition. *IEEE Transations on Audio, Speech, and Language Processing*, Vol. 15, No. 7, September 2007, pp. 2023-2032.
[18] Kinoshita, K., Delocroix, M., Nakatani, T. & Miyoshi, M. (2009). Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Transations on Audio, Speech, and Language Processing*, Vol. 17, No. 4, May 2009, pp. 534-545.
[19] Lee, A., Kawahara, T. & Shikano, K. (2001). Julius—an open source real-time large vocabulary recognition engine. *Proceedings of European Conference on Speech Communication and Technology*, September 2001, pp. 1691-1694.
[20] Maganti, H. & Matassoni, M. (2010). An audiotry modulation spectral feature for reverberant speech recognition, *Proceedings of INTERSPEECH-2010*, pp. 570-573, Makuhari, Japan, September 2010.
[21] Makino, S., Niyada, K., Mafune, Y. & Kido, K. (1992). Tohoku University and Panasonic isolated spoken word database. *Journal of the Acoustical Society of Japan*, Vol. 48, No. 12, December 1992, pp. 899-905 (in Japanese).
[22] Nakagawa, S., Hanai, K., Yamamoto, K. & Minematsu, N. (1999). Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition. *Proceedings of International Workshop on Automatic Speech Recognition and Understanding*, 1999, pp. 393-396.
[23] Nakamura, S., Hiyane, K., Asano, F. & Nishiura, T. (2000). Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, *Proceedings of IREC-2000*, pp. 965-971.
[24] Nakayama, M., Nishiura, T., Denda, Y., Kitaoka, N., Yamamoto, K., Yamada, T., Tsuge, S., Miyajima, C., Fujimoto, M., Takiguchi, T., Tamura, S., Ogawa, T., Matsuda, S., Kuroiwa, S., Takeda, K. & Nakamura, S. (2008). CENSREC-4: Development of evaluation framework for distant-talking speech recognition under reverberant environments, *Proceedings of INTERSPEECH-2008*, pp. 968-971, Brisbane, Australia, September 2008.
[25] Raut, C., Nishimoto, T. & Sagayama, S. (2006). Adaptation for long convolutional distortion by maximum likelihood based state filtering approach, *Proceedings of ICASSP-2006*, pp. 1133-1136, Toulouse, France, May 2006.
[26] Sim, B. L., Tong, Y. C. & Chang J. S. (1998). A parametric formulation of the generalized spectral subtraction method. *IEEE Transations on Speech and Audio Processing*, Vol. 6, No. 4, July 1998, pp. 328-337.
[27] Van Veen, B. & Buckley, K. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Mag.*, Vol. 5, No. 2, March 2011, pp. 4-24.