Wang et al. EURASIP Journal on Advances in Signal Processing 2012, 2012:12
http://asp.eurasipjournals.com/content/2012/1/12

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Dereverberation and denoising based on generalized spectral subtraction by multi-channel LMS algorithm using a small-scale microphone array

Longbiao Wang*, Kyohei Odani and Atsuhiko Kai

Abstract

A blind dereverberation method based on power spectral subtraction (SS) using a multi-channel least mean squares algorithm was previously proposed to suppress the reverberant speech without additive noise. The results of isolated word speech recognition experiments showed that this method achieved significant improvements over conventional cepstral mean normalization (CMN) in a reverberant environment. In this paper, we propose a blind dereverberation method based on generalized spectral subtraction (GSS), which has been shown to be effective for noise reduction, instead of power SS. Furthermore, we extend the missing feature theory (MFT), which was initially proposed to enhance the robustness of additive noise, to dereverberation. A one-stage dereverberation and denoising method based on GSS is presented to simultaneously suppress both the additive noise and nonstationary multiplicative noise (reverberation). The proposed dereverberation method based on GSS with MFT is evaluated on a large vocabulary continuous speech recognition task. When the additive noise was absent, the dereverberation method based on GSS with MFT using only 2 microphones achieves a relative word error reduction rate of 11.4 and 32.6% compared to the dereverberation method based on power SS and the conventional CMN, respectively. For the reverberant and noisy speech, the dereverberation and denoising method based on GSS achieves a relative word error reduction rate of 12.8% compared to the conventional CMN with GSS-based additive noise reduction method. We also analyze the effective factors of the compensation parameter estimation for the dereverberation method based on SS, such as the number of channels (the number of microphones), the length of reverberation to be suppressed, and the length of the utterance used for parameter estimation. The experimental results showed that the SS-based method is robust in a variety of reverberant environments for both isolated and continuous speech recognition and under various parameter estimation conditions.

Keywords: hands-free speech recognition, blind dereverberation, multi-channel least mean squares, GSS, missing feature theory

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, robust feature extraction, and model adaptation [1-3].

In this paper, 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. A blind deconvolution-based approach for the restoration of speech degraded by the acoustic environment was proposed in [4]. 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. [5,6] explored a speech dereverberation technique for which the principle was the recovery of the envelope modulations of the

* Correspondence: wang@sys.eng.shizuoka.ac.jp
Shizuoka University, Hamamatsu 432-8561, Japan

original (anechoic) speech. They applied a technique that they originally developed to treat background noise [7] to the dereverberation problem. A novel approach for multi-microphone speech dereverberation was proposed in [8]. 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 SS, in which late reverberation is treated as additive noise, was proposed in [9,10]. However, the drawback of this approach is that the optimum parameters for SS are empirically estimated from a development dataset and the late reverberation cannot be subtracted correctly as it is not modeled precisely.

In [1,11-13], an adaptive multi-channel least mean squares (MCLMS) algorithm was proposed to blindly identify the channel impulse response in a time domain. However, the estimation error of the impulse response was very large. Therefore, the isolated word recognition rate of the compensated speech using the estimated impulse response was significantly worse than that of unprocessed received distorted speech [14]. 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 variable step-size unconstrained MCLMS (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. Previously, Wang et al. [14] proposed a robust distant-talking speech recognition method based on power SS employing the MCLMS algorithm (see Figure 1a). They treated the late reverberation as additive noise, and a noise reduction technique based on power SS was proposed to estimate the power spectrum of the clean speech using an estimated power spectrum of the impulse response. To estimate the power spectra of the impulse responses, we extended the VSS-UMCLMS algorithm for identifying the impulse responses in a time domain [1] to a frequency domain. The early reverberation was normalized by CMN.

Power SS is the most commonly used SS method. A previous study has shown that GSS with a lower exponent parameter is more effective than power SS for noise reduction [15]. In this paper, instead of using power SS, GSS is employed to suppress late reverberation. We also investigate the use of missing feature theory (MFT) [16] to enhance the robustness to noise, in combination with GSS, since the reverberation cannot be suppressed completely owing to the estimation error of the impulse response. Soft-mask estimation-based MFT calculates the reliability of each spectral component from the signal-to-noise ratio (SNR). This idea is applied to reverberant speech. However, the reliability estimation is complicated in a distant-talking environment. In [17], reliability is estimated from the time lag between the power spectrum of the clean speech and that of the distorted speech. In this paper, reliability is estimated by the signal-to-reverberation ratio (SRR) since the power spectra of clean speech and the reverberation signal can be estimated by power SS or GSS using MCLMS. A diagram of the modified proposed method combining GSS with MFT is shown in Figure 1b.

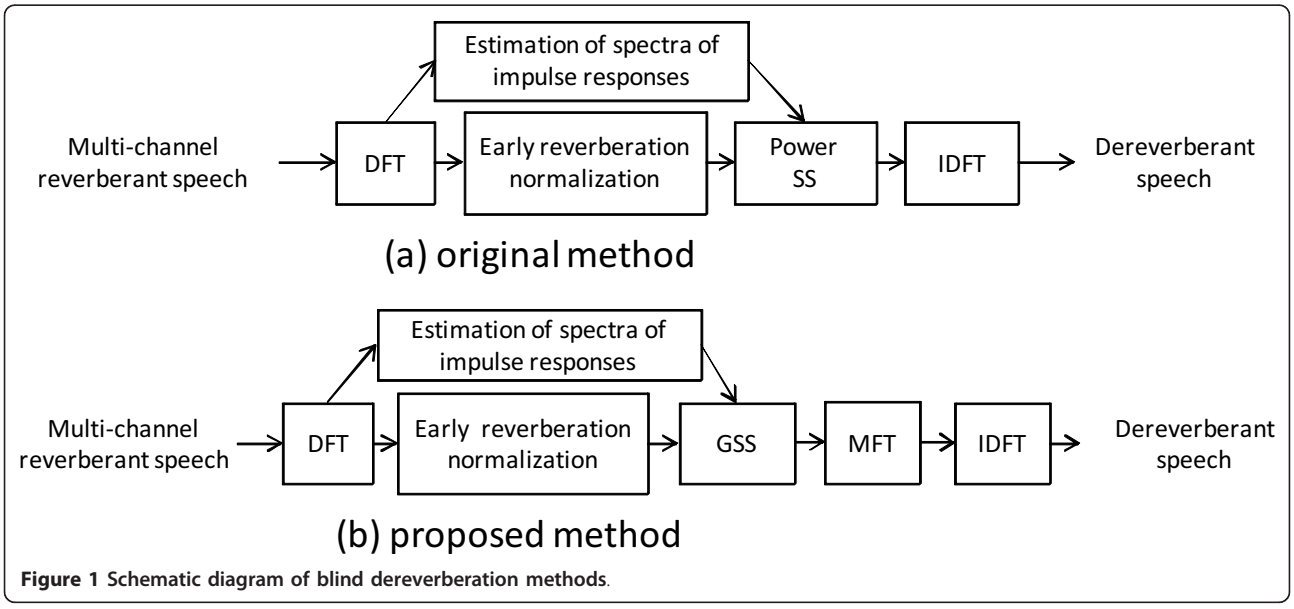

Figure 1 Schematic diagram of blind dereverberation methods.

The precision of impulse response estimation is drastically degraded when the additive noise is absent. The traditional method used two-stage processing progress, in which the reverberation suppression is performed after additive noise reduction. We present a one-stage dereverberation and denoising based on GSS. A diagram of the processing method is shown in Figure 2.

In this paper, we also investigate the robustness of the SS-based reverberation 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 SS.

The remainder of this paper is organized as follows: Section 2 describes the outline of blind dereverberation based on SS. A MFT for dereverberation is described in Section 3. A one-stage dereverberation and denoising method is proposed in Section 4, while Section 5 describes the experimental results of distant speech recognition in a reverberant environment. Finally, Section 6 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

$$x[1] = h[t] * s[t] + n[t]. \quad (1)$$

where * denotes the convolution operation. In this paper, additive noise is ignored for simplification, so Equation (1) becomes $x[t] = h[t] * s[t]$.

If the length of the impulse response is much smaller than the size T of the analysis window used for short time Fourier transform (STFT), the STFT of the distorted speech equals that of the clean speech multiplied by the STFT of the impulse response $h[t]$. However, if the length of the impulse response is much greater than the analysis window size, the STFT of the distorted speech is usually approximated by

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

where f is the frame index, $H(\omega)$ is the STFT of the impulse response, $S(f, \omega)$ is the STFT of clean speech s, D is number of reverberation windows, and $H(d, \omega)$ denotes the part of $H(\omega)$ corresponding to the frame delay d. That is, with a long impulse response, the channel distortion is no longer of a multiplicative nature in a linear spectral domain but is rather convolutional [3].

In [14], Wang et al. proposed a dereverberation method based on power SS to estimate the STFT of the clean speech $\hat{S}(f,\omega)$ based on Equation (2). The spectrum of the impulse response for the SS is blindly estimated using the method described in Section 2.3. Assuming that phases of different frames is noncorrelated for simplification, the power spectrum of Equation (2) can be approximated as

$$|X(f,\omega)|^2 \approx |S(f,\omega)|^2|H(0,\omega)|^2 + \sum_{d=1}^{D-1} |S(f-d,\omega)|^2|H(d,\omega)|^2. \quad (3)$$

The power spectrum of clean speech $|\hat{S}(f,\omega)|^2$ can be estimated as Equation (4),

$$|\hat{S}(f,\omega)|^2 = \frac{\max(|X(f,\omega)|^2 - \alpha \cdot \sum_{d=1}^{D-1} |\hat{S}(f-d,\omega)|^2|H(d,\omega|^2, \beta \cdot |X(f,\omega)|^2)}{|H(0,\omega)|^2}, \quad (4)$$

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

Furthermore, the early reverberation is compensated by subtracting the cepstral mean of the utterance. As is well known, cepstrum of the input speech x(t) is calculated as:

$$C_x = IDFT(\log(|X(\omega)|^2)) \quad (5)$$

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

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

$$|\tilde{X}(f,\omega)|^2 = |e^{DFT(C_x - \bar{C})}| = \frac{|X(f,\omega)|^2}{|\bar{X}(f,\omega)|^2}, \quad (6)$$

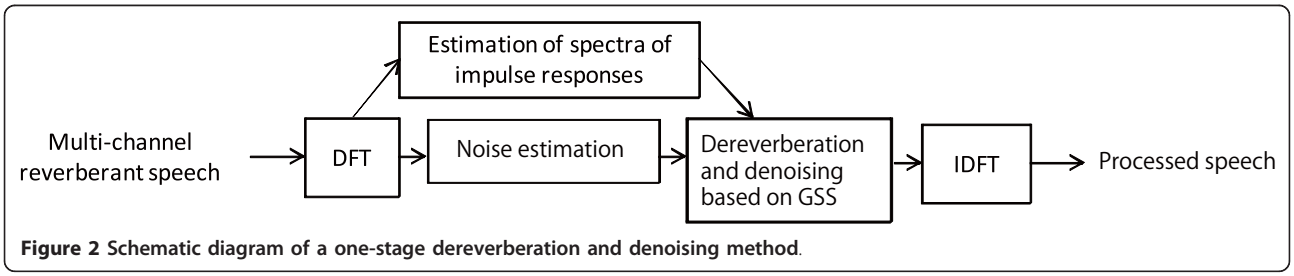

Figure 2 Schematic diagram of a one-stage dereverberation and denoising method.

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

$$\begin{aligned}|\tilde{X}(f,\omega)|^2 &= \frac{|X(f,\omega)|^2}{|\bar{X}(f,\omega)|^2} \\ &= \frac{S(f,\omega)|^2|H(0,\omega)|^2}{|\bar{X}(f,\omega)|^2} + \sum_{d=1}^{D-1}\left\{\frac{|S(f-d,\omega|^2|H(d,\omega)|^2}{|\bar{X}(f,\omega)|^2}\right\} \\ &\approx \frac{|S(f,\omega)|^2}{|\bar{S}(f,\omega)|^2} + \sum_{d=1}^{D-1}\left\{\frac{|S(f-d,\omega)|^2}{|\bar{S}(f,\omega)|^2} \times \frac{|H(d,\omega)|^2}{|H(0,\omega)|^2}\right\} \\ &= |\tilde{S}(f,\omega)|^2 + \frac{\sum_{d=1}^{D-1}\left\{|\tilde{S}(f-d,\omega)|^2 \times |H(d,\omega)|^2\right\}}{|H(0,\omega)|^2},\end{aligned} \tag{7}$$

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

$$|\tilde{S}(f,\omega)|^2 = |\tilde{X}(f,\omega)|^2 - \frac{\sum_{d=1}^{D-1}\{|\hat{S}(f-d,\omega)|^2 \times |H(d,\omega)|^2\}}{|H(0,\omega)|^2}. \tag{8}$$

The SS is used to prevent the estimated clean power spectrum being negative value; Equation (8) is adopted as:

$$|\hat{S}(f,\omega)|^2 \approx \max(|\tilde{X}(f,\omega)|^2 - \alpha \cdot \frac{\sum_{d=1}^{D-1}\{|\hat{S}(f-d,\omega)|^2|H(d,\omega)|^2\}}{|H(0,\omega)|^2}, \beta \cdot |\tilde{X}(f,\omega)|^2). \tag{9}$$

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. In this paper, we extend GSS to suppress late reverberation. Instead of the power SS-based dereverberation given in Equation (9), GSS-based dereverberation is modified as

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

where n is the exponent parameter. For power SS, the exponent parameter n is equal to 1. In this paper, the exponent parameter n is set to 0.1 as this value yielded the best results in [15].

The methods given in Eqs. (9) and (10) are referred to as *SS-based (original)* and *GSS-based (proposed) dereverberation methods*, respectively.

2.3 Compensation parameter estimation for SS by multi-channel LMS algorithm

In [1], an adaptive multi-channel LMS algorithm for blind single-input multiple-output (SIMO) system identification was proposed.

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

$$x_i * h_j = s * h_i * h_j = x_j * h_i, \ i,j = 1,2,...,N, i \neq j, \tag{11}$$

and have the following relation at time t:

$$\mathbf{x}_i^T(t)\mathbf{h}_j(t) = \mathbf{x}_j^T(t)\mathbf{h}_i(t), \quad i,j = 1,2,...,N, i \neq j, \tag{12}$$

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

$$\mathbf{x}_i(t) = [x_i(t) \quad x_i(t-1) \quad \cdots \quad x_i(t-L+1)]^T, i = 1,2,...,N,$$

where $\mathbf{x}_i(t)$ is the speech signal received from the *i-th* channel at time t and L is the number of taps of the impulse response. Multiplying Equation (12) by $\mathbf{x}_i(t)$ and taking expectation yields,

$$\mathbf{R}x_ix_i(t+1)\mathbf{h}_j(t) = \mathbf{R}x_ix_j(t+1)\mathbf{h}_i(t), i,j = 1,2,...,N, i \neq j, \tag{13}$$

where $\mathbf{R}x_ix_j(t+1) = E\{\mathbf{x}_i(t+1)\mathbf{x}_j^T(t+1)\}$. Equation (13) comprises $N(N-1)$ distinct equations. By summing up the $N-1$ cross correlations associated with one particular channel $\mathbf{h}_j(t)$, we get

$$\sum_{i=1,i\neq j}^{N} \mathbf{R}x_ix_i(t+1)\mathbf{h}_j(t) = \sum_{i=1,i\neq j}^{N} \mathbf{R}x_ix_j(t+1)\mathbf{h}_i(t), j = 1,2,...,N. \tag{14}$$

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

$$\mathbf{R}_{x+}(t+1)\mathbf{h}(t) = 0, \tag{15}$$

where

$$\mathbf{R}_{x+}(t+1) = \begin{bmatrix} \sum_{n\neq 1}\mathbf{R}x_nx_n(t+1) & -\mathbf{R}x_2x_1(t+1) & \cdots & -\mathbf{R}x_Nx_1(t+1) \\ -\mathbf{R}x_1x_2(t+1) & \sum_{n\neq 2}\mathbf{R}x_nx_n(t+1) & \cdots & -\mathbf{R}x_Nx_2(t+1) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}x_1x_N(t+1) & -\mathbf{R}x_2x_N(t+1) & \cdots & \sum_{n\neq N}\mathbf{R}x_nx_n(t+1) \end{bmatrix}, \tag{16}$$

$$\mathbf{h}(t) = [\mathbf{h}_1(t)^T \quad \mathbf{h}_2(t)^T \quad \cdots \quad \mathbf{h}_N(t)^T]^T, \tag{17}$$

$$\mathbf{h}_n(t) = [h_n(t,0) \quad h_n(t,1) \quad \cdots \quad h_n(t,L-1)]^T, \tag{18}$$

where $h_n(t, l)$ is the lth tap of the nth 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:

$$\mathbf{e}(t+1) = \tilde{\mathbf{R}}_{x+}(t+1)\hat{h}(t), \tag{19}$$

$$\tilde{\mathbf{R}}_{x+}(t+1) = \begin{bmatrix} \sum_{n\neq 1}\tilde{\mathbf{R}}x_nx_n(t+1) & -\tilde{\mathbf{R}}x_2x_1(t+1) & \cdots & -\tilde{\mathbf{R}}x_Nx_1(t+1) \\ -\tilde{\mathbf{R}}x_1x_2(x+1) & \sum_{n\neq 2}\tilde{\mathbf{R}}x_nx_n(t+1) & \cdots & -\tilde{\mathbf{R}}x_Nx_2(t+1) \\ \vdots & \vdots & \ddots & \vdots \\ -\tilde{\mathbf{R}}x_1x_N(t+1) & -\tilde{\mathbf{R}}x_2x_N(t+1) & \cdots & \sum_{n\neq N}\tilde{\mathbf{R}}x_nx_n(t+1) \end{bmatrix}, \tag{20}$$

where $\tilde{\mathbf{R}}x_ix_j(t+1) = \mathbf{x}_i(t+1)\mathbf{x}_j^T(t+1), i,j = 1,2,...,N$ 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 $t + 1$

$$J(t+1) = ||\mathbf{e}(t+1)||^2 = \mathbf{e}(t+1)^T \mathbf{e}(t+1). \quad (21)$$

By minimizing the cost function J of Equation (21), the impulse response can be blindly derived. Wang et al. [14] extended this VSS-UMCLMS algorithm [1], which identifies the multi-channel impulse responses, for processing in a frequency domain with SS applied in combination.

3. Missing feature theory for dereverberation

MFT [16] enhances the robustness of speech recognition to noise by rejecting unreliable acoustic features using a missing feature mask (MFM). The MFM is the reliability corresponding to each spectral component, with 0 and 1 being unreliable and reliable, respectively. The MFM is typically a hard and a soft mask. The hard mask applies binary reliability values of 0 or 1 to each spectral component and is generated using the signal-to-noise ratio (SNR). The reliability is 0 when the SNR is greater than a manually-defined threshold, otherwise it is 1. The soft mask is considered a better approach than the hard mask and applies a continuous value between 0 and 1 using a sigmoid function.

In a distant-talking environment, it is difficult to estimate the reliability of each spectral component since it is difficult to estimate the spectral components of clean speech and reverberant speech. Therefore, in [17], the reliability was estimated from *a priori* information by measuring the difference between the spectral components of clean speech and reverberant speech at given times. In this paper, a soft mask is calculated using the signal-to-reverberation ratio (SRR). From Equation (10), the SRR is calculated as

$$\mathrm{SRR}(f,\omega) = 10\log_{10}\left(\frac{|\hat{S}(f,\omega)|^{2n}}{\sum_{d=1}^{D-1}\left\{|\tilde{S}(f-d,\omega)|^{2n}|H(d,\omega)|^{2n}\right\}}\right). \quad (22)$$

The reliability $r(f, \omega)$ for the soft mask is generated as

$$r(f,\omega) = \frac{1}{1 + exp - a(\mathrm{SRR}(f,\omega) - b))}, \quad (23)$$

where a and b are the gradient and center of the sigmoid function, respectively, and are empirically determined. Finally, the estimated spectrum of clean speech from Equation (10) is multiplied by the reliability $r(f, \omega)$, and the inverse DFT of $|\hat{S}(f,\omega)|^{2n} r(f,\omega)$ forms the dereverberant speech.

4. One-stage dereverberation and denoising based on GSS

The precision of impulse response estimation is drastically degraded when the additive noise is present. The traditional method used two-stage processing progress, in which the reverberation suppression is performed after additive noise reduction. We present a one-stage dereverberation and denoising based on GSS. A diagram of the processing method is shown in Figure 2. At first, the spectra of additive noise and impulse responses are estimated, and then the reverberation and additive noise are suppressed simultaneously. When additive noise is present, the power spectrum of Equation (2) becomes

$$|X(f,\omega)|^2 \approx |S(f,\omega)|^2|H(0,\omega)|^2 + \sum_{d=1}^{D-1}|S(f-d,\omega)|^2|H(d,\omega)|^2 + |\bar{N}(\omega)|^2, \quad (24)$$

where $\bar{N}(\omega)$ is the mean of noise spectrum $N(\omega)$. To suppress the noise and reverberation simultaneously, Equation (10) is modified as

$$|\hat{S}(f,\omega)^{2n} \approx \max\left\{\frac{|X_N(f,\omega)|^{2n}}{\bar{X}_N(f,\omega)|^{2n}} - \alpha_1 \cdot \frac{\sum_{d=1}^{D-1}|\tilde{S}(f-d,\omega)|^{2n}|H(d,\omega)|^{2n}}{|H(0,\omega)|^{2n}}, \beta_1 \cdot \frac{|X_N(f,\omega)|^{2n}}{|\bar{X}_N(f,\omega)|^{2n}}\right\}, \quad (25)$$

$$|X_N(f,\omega)|^{2n} = \max\{|X(f,\omega)|^{2n} - \alpha_2 \cdot |\bar{N}(\omega)|^{2n}, \beta_2 \cdot |X(f,\omega)|^{2n}\}, \quad (26)$$

where $X_N(f, \omega)$ is spectrum by subtracting the spectrum of observed speech with the spectrum of noise $\bar{N}(\omega)$ and $\bar{X}_N(f,\omega)$ is mean vector of $X_N(f, \omega)$.

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. 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 [18,19] and the CENSREC-4 database [20]. Table 1 lists the details of 15 recording conditions. The illustration of microphone array is shown in Figure 3. For RWCP database, a 2-8 channel circular or linear microphone array was taken from a circular + 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

Table 1 Details of recording conditions for impulse response measurement

(a) RWCP database				
Array number	**Array type**	**Room**	**Angle**	**RT60**
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 number	**Room**	**Room size**	**RT60 (s)**	
9	Office	9.0 × 6.0 m	0.25	
10	Japanese style room	3.5 × 2.5 m	0.40	
11	Lounge	11.5 × 27.0 m	0.50	
12	Japanese style bath	1.5 × 1.0 m	0.60	
13	Living room	7.0 × 3.0 m	0.65	
14	Meeting room	7.0 × 8.5 m	0.65	
15	Elevator hall	11.5 × 6.5 m	0.75	

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

[21] 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.

Table 2 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 3 gives the conditions for SS-based dereverberation. The parameters shown in Table 3 were determined empirically. An illustration of the analysis window is shown in Figure 4. For the proposed dereverberation method based on SS, the previous clean power spectra estimated with a skip window were used to estimate the current clean power spectrum since the frame shift was half the frame length in this study [a]. The spectrum of the impulse response $H(d, \omega)$ was estimated for each utterance to be recognized. An open-source LVCSR decoder software "Julius" [22] that is based on word trigram and triphone context-dependent HMMs is used. The word accuracy for LVCSR with clean speech was 92.59% (Table 4).

5.2 Effect factor analysis of compensation parameter estimation

In this section, we describe the use of four microphones [b] to estimate the spectrum of the impulse responses without a particular explanation. Delay-and-sum beamforming (BF) was performed on the 4-channel

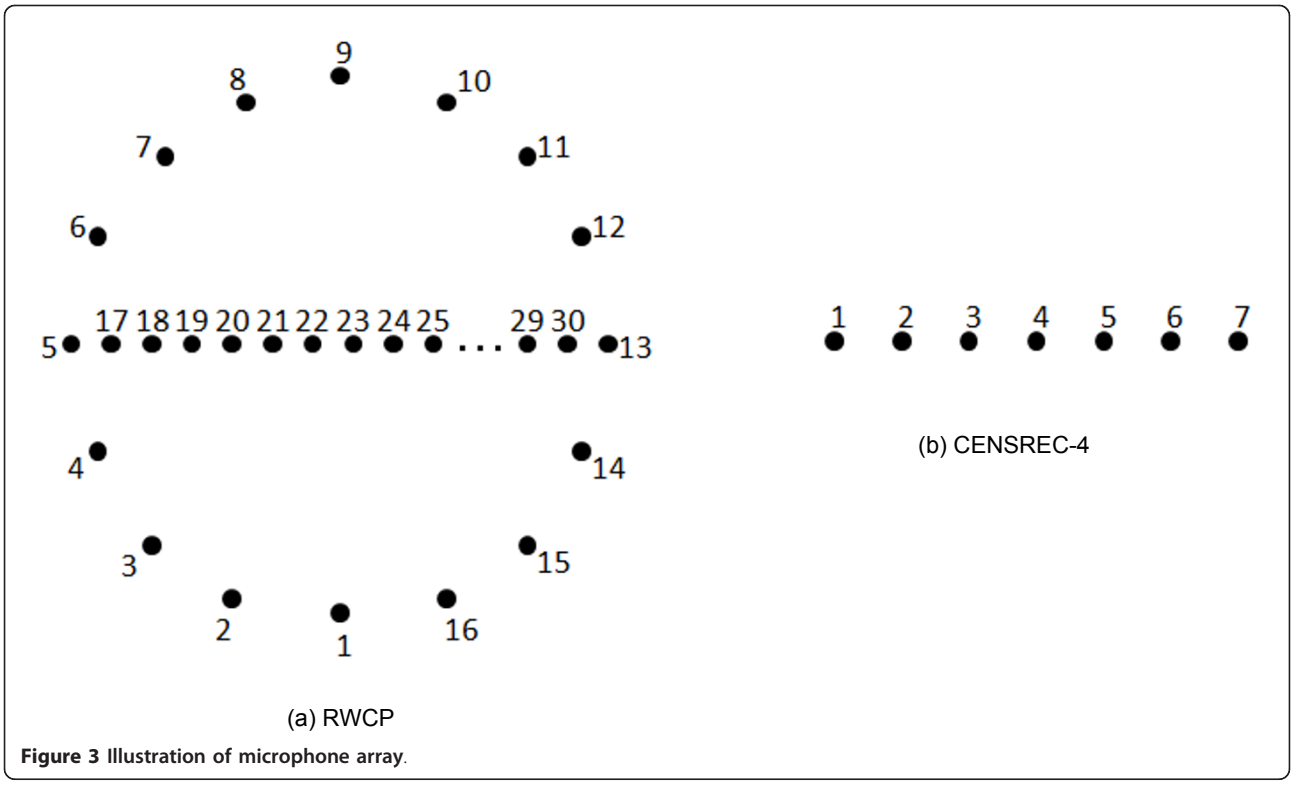

Figure 3 Illustration of microphone array.

Table 2 Conditions for speech recognition

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 (12MFCCs + Δ + Δpower)

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 SS-based dereverberation method significantly improved the speech recognition performance significantly compared with traditional CMN with beamforming [14].

In this paper, we also evaluated the SS-based dereverberation method on LVCSR with the experimental results shown in Figure 5. 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 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.

Table 3 Conditions for SS-based dereverberation

Analysis window	Hamming
Window length	32 ms
Window shift	16 ms
Number of reverberant windows D	6 (192 ms)
Noise overestimation factor α	1.0 (Power SS) 0.1 (GSS)
Spectral floor parameter β	0.15 (both)
Soft-mask gradient parameter a	0.05 (Power SS) 0.01 (GSS)
Soft-mask center parameter b	0.0 (both)

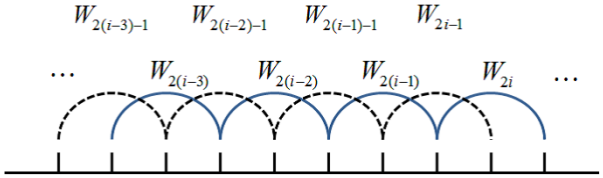

Figure 4 Illustration of the analysis window for spectral subtraction.

In this section, we also analyzed the effect factor (number of reverberation windows D in Equation (9), channel number, and length of utterance) for compensation parameter estimation for the dereverberation method based on SS using RWCP database.

The effect of the number of reverberation windows on speech recognition is shown in Figure 6. The detail results based on different number of reverberation windows D and reverberant environments (that is, different reverberation times) were shown in Table 5. The results shown on Figure 6 and Table 5 were not performed delay-and-sum beamforming. The results show 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 Figure 3a shown in Table 4 were used. The results are shown in Figure 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

Table 4 Channel number corresponding to Figure 3a using for dereverberation and denoising (RWCP database)

	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

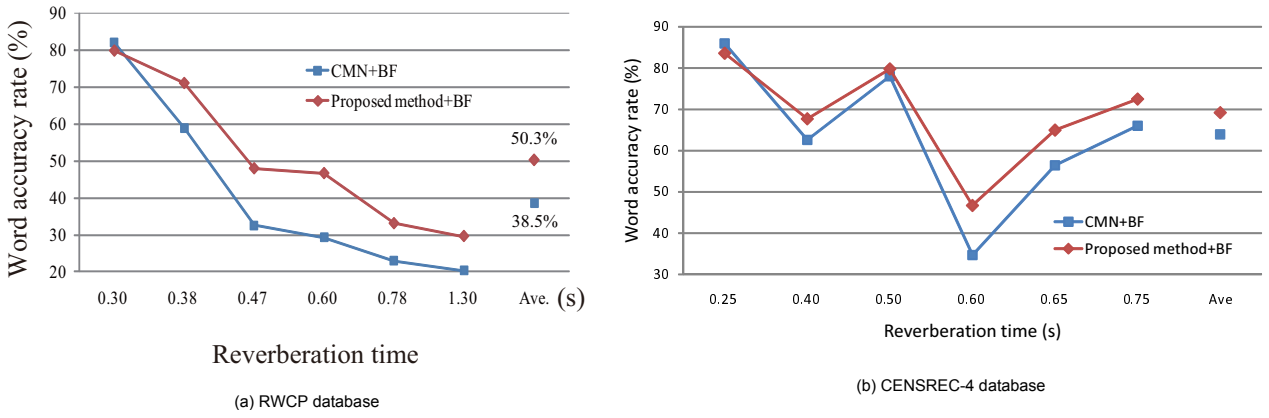

Figure 5 Word accuracy for LVCSR.

investigated, with the results shown in Figure 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.

5.3 Experimental results of dereverberation and denoising

In this section, reverberation and noise suppression using only 2 speech channels is described. [c]

In both 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 schematic of dereverberation is shown in Figure 1.

Table 6 shows the speech recognition results for the original and proposed methods. "Distorted speech #" in Table 6 corresponds to "array no" in Table 1. The word accuracy by CMN without beamforming was 40.46%. 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 Equation (9) markedly improved the speech recognition performance. The GSS-based dereverberation using Equation (10) improved speech recognition performance significantly compared with the original proposed (power SS-based dereverberation) method and CMN for

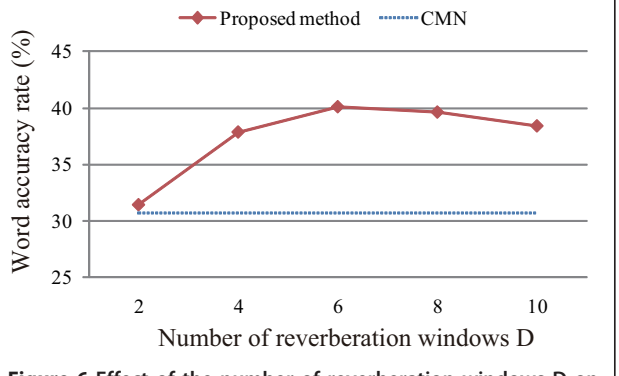

Figure 6 Effect of the number of reverberation windows D on speech recognition.

Table 5 Detail results based on different number of reverberation windows D and reverberant environments (%)

Array number #	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

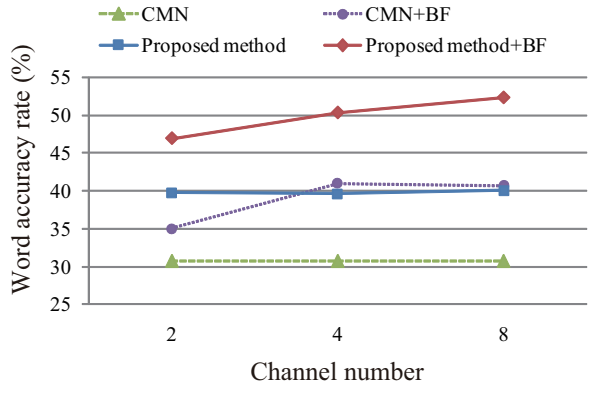

Figure 7 Effect of the number of channels on speech recognition.

all reverberant conditions. The GSS-based method without MFT 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 without MFT. When MFT was combined with both our methods, a further improvement was achieved. Finally, the GSS-based method with MFT achieved an average relative word error reduction rate of 32.6% compared to conventional CMN and 11.4% compared to the original proposed method [14].

Table 7 gives a breakdown of the word error rates obtained by the power SS- and GSS-based methods. The power SS-based method improved the substitution and deletion error rates but degraded the insertion error rate compared with CMN. The GSS-based method improved all error rates compared with the power SS-based method and achieved almost the same word insertion error as CMN.

To evaluate the proposed one-stage dereverberation and denoising based on GSS, computer room noise

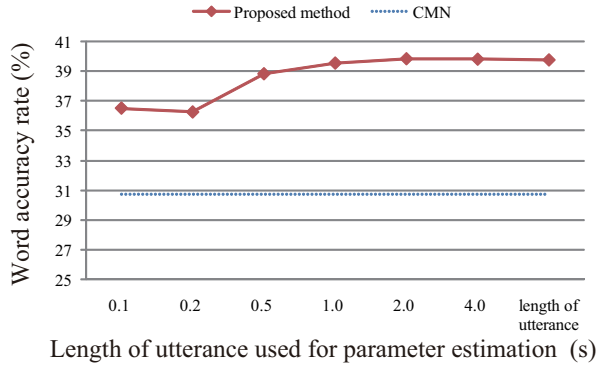

Figure 8 Effect of length of utterance used for parameter estimation on speech recognition.

Table 6 Word accuracy for LVCSR (%)

Distorted speech #	CMN only	Power SS		GSS (proposed)	
		w/o MFT	MFT	w/o MFT	MFT
2	44.35	63.34	65.15	65.95	66.47
4	27.59	40.79	44.03	49.16	47.56
5	25.61	42.55	45.75	49.29	48.31
11	73.90	79.26	78.17	80.77	80.96
12	27.06	42.28	44.91	45.38	47.83
13	29.62	50.78	54.60	56.13	58.87
15	65.24	71.67	68.31	74.35	75.93
Ave.	41.91	55.81	57.27	60.15	60.85

Delay-and-sum beamforming was performed for all methods

was added to the reverberant speech at SNRs of 15, 20, 25, and 30 dB. The noise overestimation factors α_1 and α_2 and the spectral floor parameters β_1 and β_2 in Eqs. (25) and (26) were experimentally determined as 0.07, 0.4, 0.15, and 0.1, respectively. The average results of 7 kinds of reverberant environments shown in Table 6 based on one-stage dereverberation and denoising based on GSS were shown in Table 8. The one-stage dereverberation and denoising method improved the speech recognition performance under all reverberant and noisy speech at each SNR level and reverberation time. The one-stage dereverberation and denoising method based on GSS achieved a relative word error reduction rate of 12.8% compared to the conventional CMN with GSS-based additive noise reduction method. The improvement under the additive noise condition was smaller than that for the noise-free condition. The reason might be the difference between the estimated spectrum of impulse response $H(d, \omega)$ for each condition; we compared the estimated $H(d, \omega)$ for both by first denoting the estimated spectrum of the impulse response for each as $H_1(d,\omega)$ and $H_2(d,\omega)$ and defining their average values as

$$\bar{H}_1 = \frac{\sum_{d=1}^{D} \bar{H}_1(d)}{D} = \frac{\sum_{d=1}^{D} \sqrt{\sum_{\omega} |H_1(d,\omega)|^2}}{D}, \quad (27)$$

$$\bar{H}_2 = \frac{\sum_{d=1}^{D} \bar{H}_2(d)}{D} = \frac{\sum_{d=1}^{D} \sqrt{\sum_{\omega} |H_2(d,\omega)|^2}}{D}. \quad (28)$$

Table 7 Breakdown of speech recognition errors (%)

	CMN only	Power SS		GSS (proposed)	
		w/o MFT	MFT	w/o MFT	MFT
Sub	40.61	30.48	29.37	27.39	27.42
Del	13.82	9.27	9.26	8.99	8.06
Ins	3.67	4.44	4.10	3.47	3.67

Table 8 Word accuracy for one-stage dereverberation and denoising (%)

SNR	CMN only	CMN with GSS-based noise reduction	One-stage dereverberation and denoising based on GSS
15dB	18.05	31.98	38.51
20dB	29.61	39.79	46.09
25dB	37.57	42.49	51.37
30dB	41.53	44.98	54.10
Ave.	31.69	39.81	47.52

Delay-and-sum beamforming was performed for all methods

The normalized average difference $\bar{H}_n$ between $H_1(d,\omega)$ and $H_2(d,\omega)$ is then defined as

$$\bar{H}_n = \frac{\sum_{d=1}^{D} \frac{\sum_{\omega} |H_1(d,\omega) - H_2(d,\omega)|^2}{\bar{H}_1(d)\bar{H}_2(d)}}{D}. \quad (29)$$

The average values of these estimated spectra of impulse responses and their difference are shown in Table 9. In Table 9, only the multi-channel speech of array 2 was used to calculate the average values. The result showed that $H_1(d,\omega)$ and $H_2(d,\omega)$ were quit different.

6. Conclusions

Previously, Wang et al. [14] proposed a blind dereverberation method based on power SS employing the multi-channel LMS algorithm for distant-talking speech recognition. Previous studies showed that GSS with an arbitrary exponent parameter is more effective than power SS for noise reduction. In this paper, GSS is applied instead of power SS to suppress late reverberation. However, reverberation cannot be completely suppressed owing to the estimation error of the impulse response. MFT is used to enhance the robustness of noise. Soft-mask estimation-based MFT calculates the reliability of each spectral component from SNR. In this paper, reliability was estimated through the signal-to-reverberation ratio. Furthermore, delay-and-sum beamforming was also applied to the multi-channel speech compensated by the reverberation compensation method. Our SS and GSS-based dereverberation methods were evaluated using distorted speech signals simulated by convolving multi-channel impulse responses with clean speech. When the additive noise was absent, the GSS-based method without MFT achieved an average relative word error reduction rate of 31.4% compared to conventional CMN and 9.8% compared to the power SS-based method without MFT. When MFT was combined with both our methods, further improvement was obtained. The GSS-based method with MFT achieved average relative word error reduction rates of 32.6 and 11.4% compared to conventional CMN and the original proposed method, respectively. The one-stage dereverberation and denoising method based on GSS achieved a relative word error reduction rate of 12.8% compared to the conventional CMN with GSS-based additive noise reduction method.

Table 9 Average values of the estimated spectra of impulse responses from noise-free and additive noise conditions and their difference

$\bar{H}_1$	$\bar{H}_2$	$\bar{H}_n$
0.087	0.123	0.174

In this paper, 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.

Endnotes

[a]For example, to estimate the clean power spectrum of the 2*i*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 $W_{2(i-2)}$, ... were used. [b]For RWCP database, 4 speech channels shown in Table 4 were used. For CENSREC-4 database, speech channels 1, 3, 5, and 7 shown in Figure 3b were used. [c]For RWCP database, 2 speech channels shown in Table 4 were used. For CENSREC-4 database, speech channels 1 and 3 shown in Figure 3b were used.

Competing interests
The authors declare that they have no competing interests.

Received: 15 June 2011 Accepted: 17 January 2012
Published: 17 January 2012

References

1. Y Huang, J Benesty, J Chen, *Acoustic MIMO Signal Processing* (Springer-Verlag, Berlin, 2006)
2. H Maganti, M Matassoni, An auditory based modulation spectral feature for reverberant speech recognition, in *Proceedings of INTERSPEECH 2010*, Makuhari, Japan, pp. 570–573 (2010)

3. C Raut, T Nishimoto, S Sagayama, Adaptation for long convolutional distortion by maximum likelihood based state filtering approach. Proc ICASSP. **1**, 1133–1136 (2006)
4. S Subramaniam, AP Petropulu, C Wendt, Cepstrum-based deconvolution for speech dereverberation. IEEE Trans Speech Audio Process. **4**(5), 392–396 (1996). doi:10.1109/89.536934
5. C Avendano, H Hermansky, Study on the dereverberation of speech based on temporal envelope filtering, in *Proceedings of ICSLP-1996*, Philadelphia, USA, pp. 889–892 (1996)
6. C Avendano, S Tibrewala, H Hermansky, Multiresolution channel normalization for ASR in reverberation environments, in *Proceedings of EUROSPEECH-1997*, Rhodes, Greece, pp. 1107–1110 (1997)
7. H Hermansky, EA Wan, C Avendano, Speech enhancement based on temporal processing, in *Proceedings of ICASSP-1995*, Seattle WA, USA, pp. 405–408 (1995)
8. S Gannot, M Moonen, Subspace methods for multimicrophone speech dereverberation. EURASIP J Appl Signal Process. **2003**(1), 1074–1090 (2003). doi:10.1155/S1110865703305049
9. Q Jin, Y Pan, T Schultz, Far-field speaker recognition. Proc ICASSP. **1**, 937–940 (2006)
10. Q Jin, T Schultz, A Waibel, Far-field speaker recognition. IEEE Trans ASLP. **15**(7), 2023–2032 (2007)
11. Y Huang, J Benesty, Adaptive blind channel identification: multi-channel least mean square and Newton algorithms. ICASSP II, 1637–1640 (2002)
12. Y Huang, J Benesty, Adaptive multi-channel least mean square and Newton algorithms for blind channel identification. Signal Process. **82**, 1127–1138 (2002). doi:10.1016/S0165-1684(02)00247-5
13. Y Huang, J Benesty, J Chen, Optimal step size of the adaptive multi-channel LMS algorithm for blind SIMO identification. IEEE Signal Process Lett. **12**(3), 173–175 (2005)
14. L Wang, N Kitaoka, S Nakagawa, Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm. IEICE Trans Inf Syst. **E94-D**(3), 659–667 (2011). doi:10.1587/transinf.E94.D.659
15. BL Sim, YC Tong, JS Chang, CT Tan, A parametric formulation of the generalized spectral subtraction method. IEEE Trans Speech Audio Process. **6**(4), 328–337 (1998). doi:10.1109/89.701361
16. Bhiksha Raj, Richard M Stern, Missing-feature approaches in speech recognition. IEEE Signal Process Mag. **22**(9), 101–116 (2005)
17. Kalle J Palomaki, Guy J Brown, Jon Barker, Missing data speech recognition in reverberant conditions, in *Proceedings of ICASSP-2002*, Orlando, FL, pp. 65–68 (2002)
18. http://www.slt.atr.co.jp/tnishi/DB/micarray/indexe.htm
19. S Makino, K Niyada, Y Mafune, K Kido, Tohoku University and Panasonic isolated spoken word database. J Acoust Soc Jpn. **48**(12), 899–905 (1992). (in Japanese)
20. T Nishiura, R Gruhn, S Nakamura, Evaluation framework for distant-talking speech recognition under reverberant environments, in *Proceedings of INTERSPEECH-2008*, Brisbane, Australia, pp. 968–971 (2008)
21. K Itou, M Yamamoto, K Takeda, T Takezawa, T Matsuoka, T Kobayashi, K Shikano, S Itahashi, JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. J Acoust Soc Jpn (E). **20**(3), 199–206 (1999). doi:10.1250/ast.20.199
22. A Lee, T Kawahara, K Shikano, Julius–an open source real-time large vocabulary recognition engine, in *Proceedings of European Conference on Speech Communication and Technology*, Aalborg, Denmark, pp. 1691–1694 (2001)

doi:10.1186/1687-6180-2012-12
Cite this article as: Wang *et al.*: Dereverberation and denoising based on generalized spectral subtraction by multi-channel LMS algorithm using a small-scale microphone array. *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:12.