EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation

Zhaofeng Zhang[1], Longbiao Wang[1*] and Atsuhiko Kai[2]

Abstract

Previously, a dereverberation method based on generalized spectral subtraction (GSS) using multi-channel least mean-squares (MCLMS) has been proposed. The results of speech recognition experiments showed that this method achieved a significant improvement over conventional methods. In this paper, we apply this method to distant-talking (far-field) speaker recognition. However, for far-field speech, the GSS-based dereverberation method using clean speech models degrades the speaker recognition performance. This may be because GSS-based dereverberation causes some distortion between clean speech and dereverberant speech. In this paper, we address this problem by training speaker models using dereverberant speech obtained by suppressing reverberation from arbitrary artificial reverberant speech. Furthermore, we propose an efficient computational method for a combination of the likelihood of dereverberant speech using multiple compensation parameter sets. This addresses the problem of determining optimal compensation parameters for GSS. We report the results of a speaker recognition experiment performed on large-scale far-field speech with different reverberant environments to the training environments. The proposed GSS-based dereverberation method achieves a recognition rate of 92.2%, which compares well with conventional cepstral mean normalization with delay-and-sum beamforming using a clean speech model (49.0%) and a reverberant speech model (88.4%). We also compare the proposed method with another dereverberation technique, multi-step linear prediction-based spectral subtraction (MSLP-GSS). The proposed method achieves a better recognition rate than the 90.6% of MSLP-GSS. The use of multiple compensation parameters further improves the speech recognition performance, giving our approach a recognition rate of 93.6%. We implement this method in a real environment using the optimal compensation parameters estimated from an artificial environment. The results show a recognition rate of 87.8% compared with 72.5% for delay-and-sum beamforming using a reverberant speech model.

Keywords: Hands-free speaker recognition; Blind dereverberation; Multi-channel least mean-squares; Generalized spectral subtraction; Gaussian Mixture Model

1 Introduction

Because of the existence of reverberation in far-field environments, the recognition performance for distant-talking speech/speakers is drastically degraded. The current approaches to automatic speech recognition (ASR)/speaker recognition that are robust to reverberation can be classified as speech signal processing (pre-processing), robust feature extraction, or model adaptation [1-4].

In this paper, we focus on speech signal processing for speaker identification. Beamforming is one of the simplest and most robust means of spatial filtering to suppress reverberation and background noise. This means it is able to discriminate between signals based on the physical location of their source [5]. Another general approach is cepstral mean normalization (CMN) [6,7], which has been extensively examined as a simple and effective way of reducing reverberation by normalizing the cepstral features. Because of multiple reflections and diffusions of the sound waves, the energy of previous speech is smeared over time, and overlaps with subsequent speech. This results in a duration that is much

*Correspondence: wang@vos.nagaokaut.ac.jp
[1] Nagaoka University of Technology, Nagaoka 940-2188, Japan
Full list of author information is available at the end of the article

longer than the window size of short-term spectral analysis, a problem known as late reverberation [8]. Therefore, the dereverberation of CMN is not completely effective in environments with late reverberation. Several studies have focused on mitigating the above problem [9-18]. In [9,10], a method based on mean subtraction using a long-term spectral analysis window was proposed. The result showed that subtracting the mean of the log magnitude spectrum improved ASR performance. A blind deconvolution-based approach for restoring speech that has been degraded by the acoustic environment was proposed in [19]. This scheme processed the phase-only output from two microphones using cepstrum operations and signal reconstruction theory. In [12], a multi-channel speech dereverberation method based on spectral subtraction using a statistical model to estimate the power spectrum was proposed. In the study of [13], a new set of feature parameters based on the Hilbert envelope of Gammatone filterbank outputs was proposed to reduce the effect of room reverberation in speaker recognition. A novel approach for multi-microphone speech dereverberation was proposed in [14]. The method was based on the construction of a 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 method based on multi-step linear prediction (MSLP) was proposed in [15,20]. The method first estimates late reverberations using long-term multi-step linear prediction, and then suppresses them with subsequent spectral subtraction. A reverberation compensation method for speaker recognition using spectral subtraction [16], in which late reverberation is treated as additive noise, was proposed in [18,21]. However, the drawback of this approach is that the optimum parameters for spectral subtraction are empirically estimated from a development dataset, meaning that the late reverberation cannot be subtracted correctly as it is not precisely modeled.

Previously, Wang et al. presented a distant-talking speech recognition method based on generalized spectral subtraction (GSS) employing the multi-channel least mean-squares (MCLMS) algorithm [22]. They treated late reverberation as additive noise, and proposed a noise reduction technique based on GSS [23,24] to estimate the spectrum of the clean speech using an approximated spectrum of the impulse response. To estimate the spectra of the impulse responses, a variable step-size unconstrained MCLMS algorithm for identifying the impulse responses in a time domain [1] was extended to the frequency domain. About the early reverberation, we can remove it by GSS method theoretically. But this method may cause some deviation in the MCLMS step. The estimation error of channel impulse response is inevitable, which results in unreliable estimation of power spectrum of clean speech. On the other hand, CMN is robust to reduce the channel distortion within the spectral analysis window [25]. So, early reverberation was suppressed by CMN. A speech recognition experiment showed that the GSS-based dereverberation method achieved an average relative word error reduction rate of 32.6% compared with conventional CMN with beamforming [22].

GSS-based dereverberation was applied to the field of speech recognition in a previous study [22]. However, the effect of GSS-based dereverberation on distant-talking speaker recognition is still unknown. A preliminary experiment on speaker recognition with a GSS-based method showed that dereverberation using clean speech models degraded the speaker recognition performance, but was very effective for speech recognition. This may be because the GSS-based dereverberation method causes some distortion between the speaker characteristics of clean speech and dereverberant speech. We address this problem by training speaker models using dereverberant speech obtained by suppressing early and late reverberation from arbitrary artificial reverberant speech. We assumed that the distortion of speaker characteristics in the training and test data is similar, so the GSS-based dereverberation method should be effective for speaker recognition.

It is difficult to obtain optimal compensation parameter values (that is, the noise overestimation factor α and exponent parameter n defined in Equation 5) for GSS under different conditions. We assume that the optimal compensation parameters for GSS are dependent on the acoustic environment and utterance content. A fixed compensation parameter cannot robustly suppress reverberation for all conditions. Therefore, we propose a combination of the likelihood of dereverberant speech using multiple compensation parameters for GSS. However, the computational time of this combination method is proportional to the number of compensation parameter sets. To reduce the computational cost, N speaker models with the highest likelihood are obtained using a GSS without tuning (that is, $\alpha = n = 1$). Only these N-best speaker models are used to calculate the likelihood using GSS with other compensation parameters.

With regard to speaker recognition, various models have been studied. The Gaussian mixture model (GMM) has been widely used as a speaker model [26-28]. Its use is motivated by the fact that the Gaussian components represent some general speaker-dependent spectral shapes, and by the capability of Gaussian mixtures to model arbitrary densities. Artificial neural networks [29] and support vector machines [30] have been proposed as discriminative models for the boundary between speakers. Recently, joint factor analysis and total factors [31,32] have been demonstrated as very effective mechanisms for speaker verification by compensating channel variability.

The consideration of state-of-the-art speaker models is beyond the scope of the current study. Thus, in this paper, we use GMMs for speaker identification.

The remainder of this paper is organized as follows: Section 2 describes our distant-talking speaker identification system employing a dereverberation method. The outline of blind dereverberation based on SS is described in section 3. The combination of likelihoods with various compensation parameters and its efficient computation is proposed in section 4, and section 5 describes the experimental results of distant-talking speaker recognition in a reverberant environment. Finally, section 6 summarizes the paper.

2 Distant-talking speaker recognition system employing a dereverberation method

The performance of distant-talking speech/speaker recognition is degraded remarkably by reverberation. By removing reverberation, we can expect to improve the speech/speaker recognition performance. However, very little research has studied the difference between speech recognition and speaker recognition in a distant-talking environment. For speech recognition, it is necessary to maximize the inter-phoneme variation while minimizing the intra-phoneme variation in the feature space, whereas for speaker recognition, the focus is on speaker variation instead of phoneme variation. These characteristics mean some methods that are effective in speech recognition may be not effective for speaker recognition, especially in a hands-free environment. For example, a simple and popular channel normalization method, CMN, removes both the transmission characteristics and speaker characteristics, leading to differences in the speaker recognition and speech recognition performance. A previous study [28] on distant-talking speaker recognition showed that conventional CMN gave much worse results than those without CMN, although it was very effective for speech recognition in a reverberant environment with a short reverberation time. CMN has worse speaker recognition performance than without CMN in a small reverberation environments, while the opposite is true in large-reverberation environments. This is because CMN removes the speaker characteristics, and the channel distortion (reverberation) is not very large. In the speech recognition field, GSS-based dereverberation using clean speech models showed a significant improvement [22]. However, in terms of speaker recognition, the experiment we describe in section 5 shows that it degrades the speaker recognition performance. This could be due to the GSS-based dereverberation method distorting the speaker characteristics of clean speech and dereverberant speech.

To mitigate the distortion of speaker characteristics caused by dereverberation in the test stage, we obtain dereverberant speech by suppressing early and late reverberation from arbitrary artificial reverberant speech, and use this to train the speaker models. We assume that the speaker characteristics suffer similar distortion in the training data and test data. By employing dereverberation in both the training and test stages, the transmission characteristics can be removed and the relative speaker characteristics can be maximized. Compared with speaker models trained with reverberant speech, our method is expected to exhibit a better speaker recognition performance. In previous research, GMMs trained with reverberant speech have been used for distant-talking speaker recognition. However, the mismatch of distant-talking environments between the training condition and the test condition has still not been addressed. Furthermore, when late reverberations have a large amount of energy, the performance of speech/speaker recognition cannot be improved sufficiently, even with GMMs or hidden Markov models trained with a matched reverberant condition [4,33]. This means that GMMs and hidden Markov models cannot handle severe late reverberations precisely. We can see the effect of the dereverberation step in speaker recognition in papers such as [18,21,34].

In this paper, we propose a distant-talking speaker recognition system employing a GSS-based dereverberation method. A schematic diagram of our proposed method is shown in Figure 1. In the training stage, clean speech is convoluted by arbitrary impulse responses to create artificial reverberant speech. This can reduce the experimental cost, because real reverberant speech is not necessary. We introduce GSS-based dereverberation in section 3. This is performed to suppress both early and late reverberations. Finally, the dereverberant speech is used to train speaker models. In the test stage, the reverberation of multi-channel distorted speech (artificial reverberant speech or real reverberant speech) is removed by the GSS-based dereverberation method, and then the dereverberant speech is used to perform distant-talking speaker recognition.

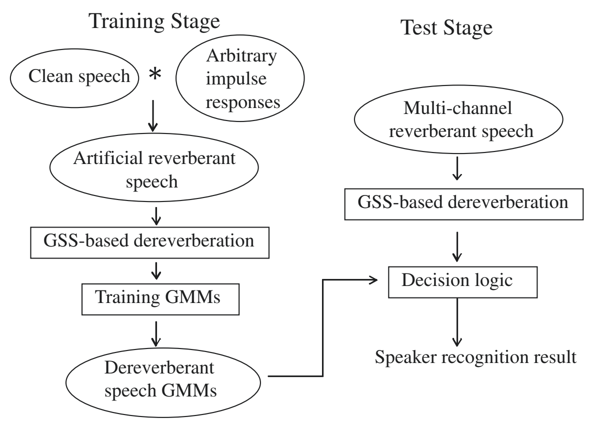

Training Stage
Test Stage
Clean speech
*
Arbitrary impulse responses
Artificial reverberant speech
GSS-based dereverberation
Training GMMs
Dereverberant speech GMMs
Multi-channel reverberant speech
GSS-based dereverberation
Decision logic
Speaker recognition result

Figure 1 Schematic diagram of distant-talking speaker recognition system.

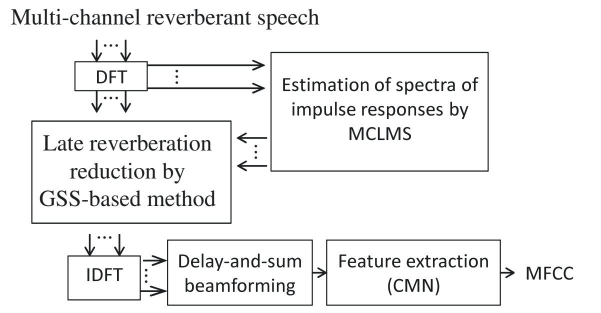

Multi-channel reverberant speech
DFT
Estimation of spectra of impulse responses by MCLMS
Late reverberation reduction by GSS-based method
IDFT
Delay-and-sum beamforming
Feature extraction (CMN)
MFCC

Figure 2 Schematic diagram of GSS-MCLMS-based dereverberation method.

3 Outline of blind dereverberation

3.1 Dereverberation based on GSS

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

$$x[t] = h[t] * s[t]\,, \tag{1}$$

where * denotes the convolution operation. If the length of the impulse response is much smaller than the size T of the analysis window used for the discrete-time Fourier transform (DTFT), the DTFT of the distorted speech equals that of the clean speech multiplied by the DTFT of the impulse response $h[\,t]$. However, if the length of the impulse response is much greater than the analysis window size, the DTFT of the distorted speech is usually approximated by

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

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, 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 instead convolutional.

In [22], Wang et al. proposed a dereverberation method based on GSS to estimate the STFT of the clean speech $\hat{S}(f,\omega)$ based on Equation 2. Assuming that phases of different frames are noncorrelated for simplification, the power spectrum of Equation 2 can be approximated as Equation 3:

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

The power spectrum $|\hat{X}(f,\omega)|^2$ obtained by reducing the late reverberation can be estimated as

$$\begin{aligned} |\hat{X}(f,\omega)|^2 &= |\hat{S}(f,\omega)|^2|\hat{H}(0,\omega)|^2 \\ &\approx \max\Bigg\{|X(f,\omega)|^2 \\ &\quad - \alpha \cdot \sum_{d=1}^{D-1}\left\{|\hat{S}(f-d,\omega)|^2|\hat{H}(d,\omega)|^2\right\}, \\ &\quad \beta \cdot |X(f,\omega)|^2\Bigg\} \\ &= \max\Bigg\{|X(f,\omega)|^2 \\ &\quad - \alpha \cdot \frac{\sum_{d=1}^{D-1}\left\{|\hat{X}(f-d,\omega)|^2|\hat{H}(d,\omega)|^2\right\}}{|\hat{H}(0,\omega)|^2}, \\ &\quad \beta \cdot |X(f,\omega)|^2\Bigg\}, \end{aligned} \tag{4}$$

where α is the noise overestimation factor, β is the spectral floor parameter for avoiding negative or underflow values, $|\hat{S}(f,\omega)|^2$ is the power spectrum of estimated clean speech, and $\hat{H}(d,\omega), d = 0, 1 \ldots D-1$ is the STFT of the

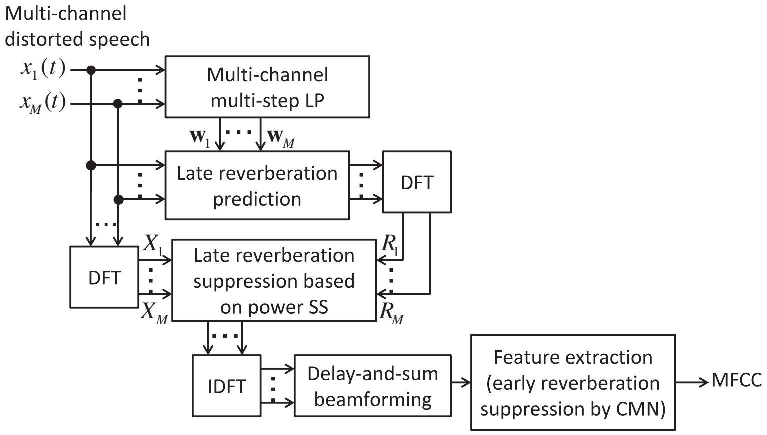

Multi-channel distorted speech
$x_1(t)$
$x_M(t)$
Multi-channel multi-step LP
$\mathbf{w}_1$
$\mathbf{w}_M$
Late reverberation prediction
DFT
DFT
X_1
X_M
Late reverberation suppression based on power SS
R_1
R_M
IDFT
Delay-and-sum beamforming
Feature extraction (early reverberation suppression by CMN)
MFCC

Figure 3 Schematic diagram of MSLP-based dereverberation method.

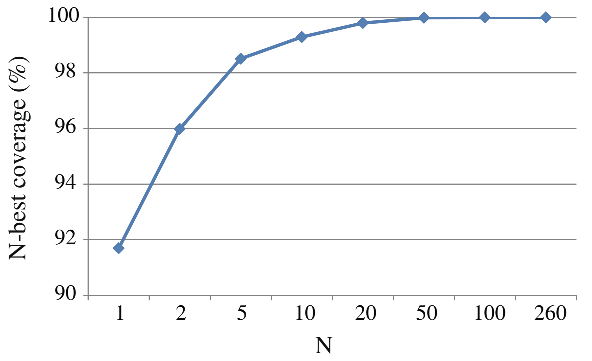


Figure 4 N-best coverage of distant-talking speaker recognition.

impulse response, which can be blindly estimated by the MCLMS algorithm method mentioned in [22]. D is the number of reverberation windows.

Previous studies have shown that GSS with an arbitrary exponent parameter is more effective than power SS for noise reduction [23,24]. In this paper, GSS is used to suppress late reverberation, and early reverberation is compensated by subtracting the cepstral mean of the utterance at the feature extraction stage.

The spectrum $|\hat{X}(f,\omega)|^{2n}$ obtained by reducing the late reverberation can be estimated as

$$|\hat{X}(f,\omega)|^{2n} \approx \max\left\{|X(f,\omega)|^{2n} - \alpha\cdot\frac{\sum_{d=1}^{D-1}\left\{|\hat{X}(f-d,\omega)|^{2n}|\hat{H}(d,\omega)|^{2n}\right\}}{|\hat{H}(0,\omega)|^{2n}},\ \beta\cdot|X(f,\omega)|^{2n}\right\}, \tag{5}$$

where $|\hat{X}(f,\omega)|^{2n} = |\hat{S}(f,\omega)|^{2n}|\hat{H}(0,\omega)|^{2n}$, $|\hat{S}(f,\omega)|^{2n}$ is the spectrum of estimated clean speech and n is the exponent parameter. When $n = 1$, Equation 5 is a power spectral subtraction-based method.

A schematic diagram of our proposed GSS-based dereverberation method is shown in Figure 2. It uses the spectra of impulse responses, which are estimated by MCLMS, to reduce the late reverberation in reverberant speech.

The spectrum of dereverberant speech is then inverted into the time domain, and delay-and-sum beamforming [a] is performed on the multi-channel speech. Finally, the early reverberation is normalized by CMN at the feature extraction stage.

3.2 Compensation parameter estimation for GSS by MCLMS

In [1,35-37], an adaptive MCLMS algorithm for blind single-input multiple-output (SIMO) system identification was proposed.

A variable step-size unconstrained MCLMS (VSS-UMCLMS) algorithm was proposed to minimize the cost function J in the time-domain [37]. Wang et al. [38] extended the time-domain VSS-UMCLMS algorithm to the frequency domain to estimate the compensation parameters for GSS-based dereverberation.

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

$$X_i * H_j = S * H_i * H_j = X_j * H_i, \quad i,j = 1,2,\ldots,N, i \neq j, \tag{6}$$

and have the following relation at frequency ω of frame d:

$$\mathbf{X}_i^T(d)\mathbf{H}_j(d) = \mathbf{X}_j^T(d)\mathbf{H}_i(d), \quad i,j = 1,2,\ldots,N, i \neq j, \tag{7}$$

where $\mathbf{H}_i(d)$ is the ith impulse response at frame index f and

$$\mathbf{X}_i(d) = \left[X_i(d)\ \ X_i(d-1)\ \ \cdots\ \ X_i(d-D+1)\right]^T,\quad i = 1,2,\ldots,N,$$

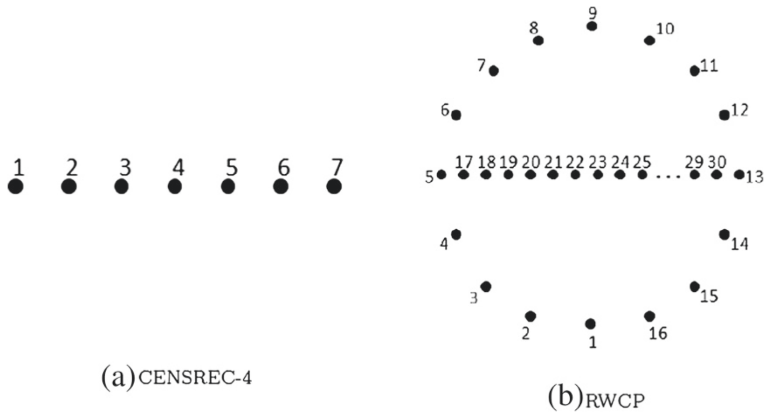
where $\mathbf{X}_i(d)$ is the speech signal received from the ith channel at frame d and D is the number of frames of the impulse response. Multiplying Equation 7 by $\mathbf{X}_i(d)$ and taking the expectation yields

$$\mathbf{R}_{X_iX_i}(d)\mathbf{H}_j(d) = \mathbf{R}_{X_iX_j}(d)\mathbf{H}_i(d), i,j = 1,2,\ldots,N, i \neq j, \tag{8}$$

Table 1 Details of recording conditions for impulse response measurement

	Array no	Room	Array type	RT60 (s)	SRR
CENSREC-4 database for training	1	Japanese style room	Linear	0.40	3.24
	2	Japanese style bath	Linear	0.60	3.28
	3	Elevator hall	Linear	0.75	2.98
RWCP database for test	4	Echo room (cylinder)	Circle	0.38	3.45
	5	Tatami-floored room (S)	Circle	0.47	2.89
	6	Tatami-floored room (L)	Circle	0.60	3.12
	7	Conference room	Circle	0.78	3.30
	8	Echo room (panel)	Linear	1.30	2.88

RT60 (s), reverberation time in room; S, small; L, large; SRR, signal-to-reverberation ratio [42], calculated from artificial reverberant data.



(a) CENSREC-4

(b) RWCP

Figure 5 Illustration of microphone array. (a) CENSREC-4. **(b)** RWCP.

where $\mathbf{R}_{X_iX_j}(d) = E\{\mathbf{X}_i(d)\mathbf{x}_j^T(d)\}$. Equation 8 comprises $N(N-1)$ distinct equations. By summing the $N-1$ cross-correlations associated with one particular channel $\mathbf{H}_j(d)$, we get

$$\sum_{i=1,i\neq j}^{N} \mathbf{R}_{X_iX_i}(d)\mathbf{H}_j(d) = \sum_{i=1,i\neq j}^{N} \mathbf{R}_{X_iX_j}(d)\mathbf{H}_i(d), \quad j = 1, 2, \ldots, N. \tag{9}$$

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

$$\mathbf{R}_{X+}(d)\mathbf{H}(d) = 0, \tag{10}$$

where

$$\mathbf{R}_{X+}(d) = \begin{bmatrix} \sum_{n\neq 1}\mathbf{R}_{X_nX_n}(d) & -\mathbf{R}_{X_2X_1}(d) & \cdots & -\mathbf{R}_{X_NX_1}(d) \\ -\mathbf{R}_{X_1X_2}(d) & \sum_{n\neq 2}\mathbf{R}_{X_nX_n}(d) & \cdots & -\mathbf{R}_{X_NX_2}(d) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{X_1X_N}(d) & -\mathbf{R}_{X_2X_N}(d) & \cdots & \sum_{n\neq N}\mathbf{R}_{X_nX_n}(d) \end{bmatrix}, \tag{11}$$

$$\mathbf{H}(d) = \left[\mathbf{H}_1(d)^T \quad \mathbf{H}_2(d)^T \quad \cdots \quad \mathbf{H}_N(d)^T\right]^T, \tag{12}$$

$$\mathbf{H}_n(d) = [H_n(d,0) \quad h_n(d,1) \quad \cdots \quad H_n(d,D-1)]^T, \tag{13}$$

where $H_n(d,l)$ is the lth frame of the nth impulse response at correspond frame d. 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 estimated channel impulse responses deviate from the true value, the error vector at frame d is produced by:

$$\mathbf{e}(d) = \tilde{\mathbf{R}}_{X+}(d)\hat{\mathbf{H}}(d), \tag{14}$$

$$\tilde{\mathbf{R}}_{X+}(d) = \begin{bmatrix} \sum_{n\neq 1}\tilde{\mathbf{R}}_{X_nX_n}(d) & -\tilde{\mathbf{R}}_{X_2X_1}(d) & \cdots & -\tilde{\mathbf{R}}_{X_NX_1}(d) \\ -\tilde{\mathbf{R}}_{X_1X_2}(d) & \sum_{n\neq 2}\tilde{\mathbf{R}}_{X_nX_n}(d) & \cdots & -\tilde{\mathbf{R}}_{X_NX_2}(d) \\ \vdots & \vdots & \ddots & \vdots \\ -\tilde{\mathbf{R}}_{X_1X_N}(d) & -\tilde{\mathbf{R}}_{X_2X_N}(d) & \cdots & \sum_{n\neq N}\tilde{\mathbf{R}}_{X_nX_n}(d) \end{bmatrix}, \tag{15}$$

where $\tilde{\mathbf{R}}_{X_iX_j}(d) = \mathbf{X}_i(d)\mathbf{X}_j^T(d), i,j = 1, 2, \ldots, N$ and $\hat{\mathbf{H}}(d)$ is the estimated model filter at frame d. Here, the tilde in $\tilde{\mathbf{R}}_{X_iX_j}$ distinguishes this instantaneous value from its mathematical expectation $\mathbf{R}_{X_iX_j}$.

This error can be used to define a cost function at frame d

$$J(d) = \|\mathbf{e}(d)\|^2 = \mathbf{e}(d)^T\mathbf{e}(d). \tag{16}$$

By minimizing the cost function J in Equation 16, the impulse response can be blindly derived.

3.3 Dereverberation method based on multiple-step linear prediction

In [15], MSLP was implemented for our reverberation calculation. Linear prediction is a method of generating an inverse filter through a prediction coefficient, which is an effective means of estimating the inverse system. In particular, multi-channel linear prediction can estimate the inverse filter blindly. For comparison with our proposed method, we introduce this alternative approach. A schematic diagram of MSLP is shown in Figure 3.

Table 2 Channel numbers corresponding to Figure 5 using for dereverberation

	Linear array	Circle array
CENSREC-4	1, 3, 5, 7	—
RWCP	17, 21, 25, 29	1, 5, 9, 13

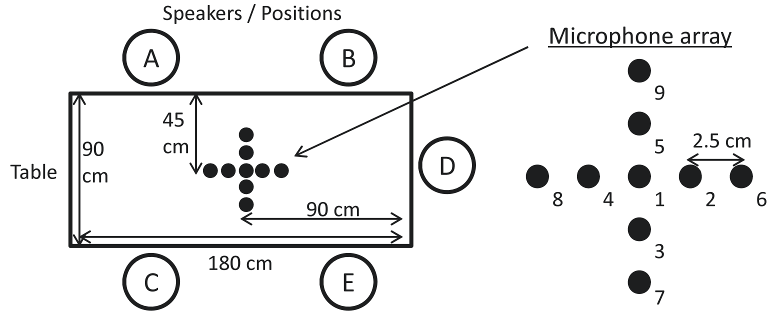


Figure 6 Illustration of recording settings and microphone array.

The reverberant speech $x(t)$ is

$$x[t] = h[t] * s[t] = \sum_{i=0}^{K-1} s(t-i)h(i), \tag{17}$$

where K is the length of the impulse response. Considering the step size D of early reverberation, Equation 17 can be rewritten as

$$x_m[t] = \sum_{k=0}^{D-1} h_m(k)s(t-k) + \sum_{k=D}^{K-1} h_m(k)s(t-k), \tag{18}$$

where $x_m(t)$ is the observed signal from the mth microphone. The first part of the right-hand side of this equation is the early reverberation, and the second part is the late reverberation. Using the MSLP method, we have

$$x_m[t] = \sum_{k=i}^{M} \sum_{k=0}^{L-1} w_{m,i}(k)x_i(t-D-k) + d_m(t), \tag{19}$$

where L is the linear prediction order and $w_{m,i}$ is the prediction coefficient. When $D = 1$, we have multi-channel linear prediction. To calculate the appropriate $w_{m,i}$, the present signal of the mth microphone $x_m(t)$ should be presented as the sum of the weighted signals of the previous D samples (first term of Equation 19) and signal $d_m(t)$ without late reverberation (second term of Equation 19).

After the optimization of $w_{m,i}$, the dereverberant speech can be calculated by the SS method. In [15], the $w_{m,i}$ are calculated by minimizing the mean square energy of the prediction residual.

4 Combination method and its efficient computation

It is difficult to determine the optimum exponent parameter n and the noise overestimation factor α for GSS. In this study, we use a combination of the various speaker model likelihoods with different compensation parameter sets.

When a combination of multiple methods is used to identify the speaker, the likelihood of speaker models with different compensation parameter sets is linearly coupled to produce a new score L^k_{comb}, given by:

$$L^k_{\text{comb}} = \frac{1}{I} \sum_{i=1}^{I} L^k_i, \quad k = 1, 2, \cdots, K, \tag{20}$$

where L^k_i is the likelihood produced by the kth speaker model with the ith compensation parameter set. K is the number of registered speakers and I denotes the number of compensation parameter sets. The speaker with the maximum likelihood is determined as the target speaker. As a result of this procedure, special tuning is not necessary for GSS.

However, the computational time increases linearly according to the number of compensation parameter sets. In this study, an efficient computational method is proposed. Coverage of the N-best speaker recognitions is illustrated in Figure 4[b]. The number of target speakers is 260. The result shows that the coverage is over 99% for the 10-best likelihoods , and almost 100% for the 50-best likelihoods, even in a distant-talking environment. That is, there is no need to calculate the likelihood of all speaker models in the combination stage. The efficient computational method can be summarized as follows: Initially, the power SS (that is, compensation parameter $n = 1$) is used to suppress the reverberation, and the likelihoods of all speaker models are calculated. Second, the speaker models with the top N-best likelihoods are used to calculate a new likelihood according to different compensation parameter sets. Finally, the likelihood calculated by

Table 3 Conditions for speaker recognition

Sampling frequency	16 kHz
Frame length	25 ms
Frame shift	10 ms
Feature space	25 dimensions with CMN (12 MFCCs + Δ + Δpower)
Acoustic model	GMMs with 128 diagonal covariance matrices

Table 4 Conditions for GSS-based dereverberation

Analysis window	Hamming
Window length	32 ms
Window shift	16 ms
Number of reverberant windows D	6 (192 ms)
Spectral floor parameter β	0.15

a different compensation parameter set is combined to determine the target speaker.

In our previous work [22], the speech recognition performances using DTFT of impulse response estimated by MCLMS with each sentence and impulse response condition were almost same. So in this paper, each impulse response condition used the same impulse. The total computational time T_A for speaker identification is about $\frac{T_M}{s} + T_F + T_L$, where T_F and T_L are the computational times for the feature extraction and likelihood calculation of K speaker models. T_M is the time for the MCLMS algorithm. As we run the MCLMS algorithm for each reverberation condition only once, the time of T_M for a single speech is $\frac{T_M}{s}$, where s is the number of test sets. Because our experiment uses a large number of test sets, the value of $\frac{T_M}{s}$ is very small, and can be neglected here. The computational time for the combination (that is, conventional combination method) of various results with I parameter sets is $T_A^{\text{comb}} = I(T_F + T_L) = IT_A$. The computational time for our proposed efficient combination method using the N-best likelihoods is

$$\begin{aligned} T_E^{\text{comb}} &= T_F + T_L + (I-1)T_F + \frac{(I-1)N}{K} T_L \\ &= T_A + \frac{1}{\gamma+1}\left(\frac{(I-1)N}{K}\gamma + I - 1\right)T_A, \end{aligned} \tag{21}$$

where T_L equals γT_F. The computational cost has therefore been decreased compared with the conventional combination method.

5 Experiments

5.1 Experimental setup

Firstly, the proposed method for hands-free speaker identification was evaluated using artificial reverberant speech for determining the most suitable parameters. Then we implemented the method for real reverberant speech with suited parameters[c].

In order to compare our work with other dereverberation method. We compared the performance of our proposed method and multi-step linear prediction [15] (MSLP) both in artificial and real reverberant environment.

Eight multi-channel impulse responses were selected from the Real World Computing Partnership (RWCP) sound scene database [39] and the CENSREC-4 database [40]. These were convoluted with clean speech to create artificial reverberant speech. A large-scale database, the Japanese Newspaper Article Sentence (JNAS) [41] corpus, was used as clean speech. The utterances in the training data were composed of 130 male and female speakers, with 10 utterances taken from each. Each speaker gave 20 utterances for the test data. The average time for all utterances was about 5.8 s.

Table 1 lists the impulse responses for the training and test sets. The illustration of microphone array is shown in Figure 5. Channel numbers corresponding to Figure 5 using for dereverberation shown in Table 2 were used. For the RWCP database, a four-channel circular or linear microphone array was taken from a circular + linear microphone array (30 channels). The circular array had a diameter of 30 cm. The microphones in 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, four-channel microphones were taken from a linear microphone array (seven channels), with the microphones located at 2.125-cm intervals. Impulse responses were measured at several positions 0.5 m from the microphone array.

We also use reverberant speech from a real environment in our experiment. The speech was collected in a meeting room of size 7.7 m × 3.3 m × 2.5 m ($D \times W \times H$). The utterances were collected from 20 male speakers. Each speaker made 9 training utterances. In total, 400 test utterances were recorded. Speakers were seated on chairs (labeled A to E in Figure 6), and were recorded by a multi-channel recording device. The heights of the

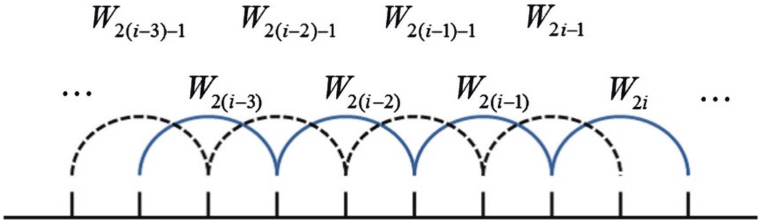


Figure 7 Illustration of the analysis window for spectral subtraction.

microphone array and the utterance position of each speaker were about 0.8 and 1.0 m, respectively. We used a nine-channel microphone array (Figure 6), and collected the test data using distant microphone arrays for four channels of microphones 6, 7, 8, and 9. A pin microphone recorded speech in the distant-talking and close-talking environments. The training data were collected by a close microphone, and the CENSREC-4 database (CENSREC-4 impulse response) was used to produce artificial reverberant speech.

Table 3 gives the conditions for speaker identification. We used 25-dimensional mel-frequency cepstral coefficients (MFCCs) and GMMs with 128 mixtures. Table 4 gives the conditions for GSS-based dereverberation (the same for MCLMS- and MSLP-based methods). The parameters shown in Table 4 were determined empirically. An illustration of the analysis window is shown in Figure 7. For the proposed dereverberation method based on spectral subtraction, the previous clean spectra estimated with a skip window were used to estimate the current clean spectrum since the frame shift was half the frame length in this study [d]. The spectrum of the impulse response $H(d, \omega)$ was estimated for each utterance to be recognized.

This study compares five methods. A description of each methods is presented in Table 5. For each method, we performed CMN with delay-and-sum beamforming. Clean speech models, which were directly trained by clean speech, were used as speaker models for *method 1* and *method 2*. For *method 1*, only CMN with beamforming was used to reduce the reverberation. The GSS-MCLMS based dereverberation was performed at the test stage for *method 2*, which is the same as the condition for hands-free speech recognition [22]. Reverberant speech models, which were trained using artificial reverberant speech with three types of CENSREC-4 impulse responses (see Table 1a), were used as speaker models for *method 3*. *Method 5* is our proposed method. For this, the reverberation in both the training and test data was suppressed by MCLMC-GSS based dereverberation, and the dereverberant speech was used to train dereverberant speech GMMs. For comparison, we also used an existing MSLP-GSS as *Method 4* with dereverberant speech in both the training and test data.

Table 5 Description of each speaker recognition method

Method number	Speaker models	Processing at test stage
1 (Baseline)	Clean speech models	CMN with beamforming
2 (Method in [22])	Clean speech models	GSS-based dereverberation
3	Reverberant speech models	CMN with beamforming
4 (MSLP-based method)	Dereverberant speech models based on MSLP-GSS	MSLP-GSS-based dereverberation
5 (Proposed method)	Dereverberant speech models based on MCLMS-GSS	MCLMS-GSS-based dereverberation

Table 6 Distant-talking speaker recognition rates of artificial data (%)

Method number	Number of impulse response condition for test					Average
	4	5	6	7	8	
1	66.7	53.3	43.2	43.7	38.3	49.0
2	53.1	32.9	25.6	25.3	29.1	33.2
3	91.6	88.4	86.5	87.6	88.0	88.4
4	93.7	90.2	89.8	89.9	89.2	90.6
5	94.0	90.6	91.0	90.5	92.3	91.7

5.2 Experimental results

5.2.1 Experimental results of artificial reverberant speech

The hands-free speaker identification results for the five methods are compared in Table 6. 'Number of impulse response conditions for test' in Table 6 denotes the 'Array no.' in Table 1b. In previous research, the speech recognition results for reverberant environments with clean

Table 7 Comparison of results of artificial data with different compensation parameter sets and combination methods for speaker identification

	Number of impulse response condition for test (%)					Average (%)
	4	5	6	7	8	
Parameters (n, α)						
(0.1, 0.1)	95.2	90.2	87.2	90.5	92.6	91.1
(0.3, 0.3)	*96.2*	89.9	89.6	87.8	89.9	90.7
(0.5, 0.5)	*96.2*	*91.1*	91.4	90.1	92.3	*92.2*
(0.7, 0.7)	95.0	90.3	91.4	*90.8*	92.5	92.0
(1.0, 1.0)	94.0	90.6	91.0	90.5	92.3	91.7
(0.1, 0.2)	94.6	88.9	88.0	86.0	90.4	89.6
(0.3, 0.6)	95.8	89.9	89.8	88.8	91.3	91.1
(0.5, 1.0)	95.3	91.0	91.0	90.5	*93.1*	*92.2*
(0.7, 1.4)	94.5	90.9	*91.7*	90.6	92.9	92.1
(1.0, 2.0)	93.8	89.8	90.9	90.3	92.0	91.3
Conventional combination	96.2	92.5	92.6	92.5	94.1	93.6
Efficient combination ($N = 5$)	96.2	92.5	92.6	92.5	94.1	93.6

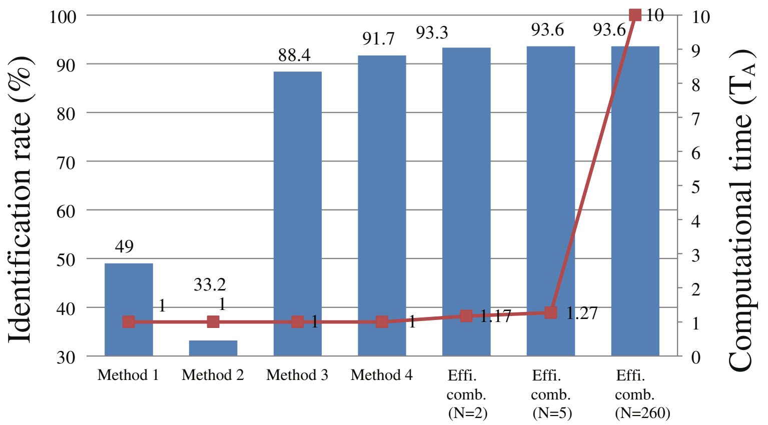


Figure 8 Comparison of distant-talking speaker recognition performance and computational cost. Methods 1 to 4 are described in Table 5. 'Effi. comb.' denotes the 'Efficient combination' described in section 4.

speech models improved when using the GSS-based dereverberation method [22]. However, *method 2* proposed in [22] degraded the speaker identification performance in the speaker identification field. *Method 3*, which was based on reverberant speech models, improved speaker recognition significantly because multiple reverberant environments were trained. However, the reverberation was not suppressed, so employing blind dereverberation may give a further improvement. The proposed method without parameter tuning (that is, $\alpha = n = 1$), which suppressed the reverberation in both training and test data, outperformed all the other methods under all reverberant environments. The proposed method achieved a relative error reduction of 83.7% compared with the baseline (*method 1*) and 28.4% compared with reverberant speech models (*method 3*). Furthermore, the proposed method performed better than the existing *method 4* with a relative error reduction of 11.7%.

The performance of the proposed GSS-based dereverberation method may vary with different compensation parameters. We confirmed this and compared the performance of the proposed method with different parameters (noise overestimation α and exponent parameter n). The results are given in Table 7. For GSS, the exponent parameter n is often set in the range 0.1 to 1 [23,24]. Thus, in this study, the exponent parameter n was set as 0.1, 0.3, 0.5, 0.7, and 1.0, and the noise overestimation factor α was set as $\alpha = n$ or $\alpha = 2n$. The results show that the optimum parameter depends on the reverberant environment, and is very difficult to determine. By combining the results with various compensation parameter sets, we achieved a relative error reduction of 17.9% compared with the individual results with the optimum parameter. The GSS parameter determination increased the computational cost. For the conventional combination method, the computational time T_A^{comb} is 10 (the number of parameter sets I is 10) times the computational time for the individual method T_A. The computational time

Table 8 Comparison of results of artificial data with different parameter of β and combination methods for speaker identification

Parameter (β)	Number of impulse response condition for test (%)					Average (%)
	4	5	6	7	8	
(0.05)	92.4	85.2	84.6	85.8	90.5	87.7
(0.10)	95.0	90.1	89.9	89.0	89.6	90.7
(0.15)	*96.2*	*91.1*	*91.4*	*90.1*	*92.3*	*92.2*
(0.20)	95.2	89.5	90.2	89.9	90.5	91.1
(0.25)	93.5	90.2	87.7	88.0	91.2	90.1

Table 9 Speaker recognition rates in real environment

Method number	Speaker models	Processing at test stage	Recognition rate (%)
1 (Baseline)	Clean speech models	CMN with beamforming	61.5
2	Reverberant speech models	CMN with beamforming	72.5
3 (LTLSS-based method)	Dereverberant speech models based on LTLSS	LTLSS-based dereverberation	81.0
4 (MSLP-GSS-based method)	Dereverberant speech models based on MSLP-GSS	MSLP-GSS-based dereverberation	83.8
5 (MCLMS-GSS-based method)	Dereverberant speech models based on MCLMS-GSS	MCLMS-GSS-based dereverberation	87.8

T_E^{comb} for our proposed efficient combination method is 1.27 T_A[e], and about 1/8 T_A^{comb} when the performance is the same as the conventional combination method, which uses the likelihoods of all the speaker models. As a result, the proposed efficient combination method achieved a relative error reduction of 87.5% compared with the baseline, and 44.8% compared with reverberant speech models, for almost the same computational cost. A comparison of the performance and computational cost of the proposed efficient combination method using the 2-best likelihoods, 5-best likelihoods, and 260-best likelihoods (that is, the conventional combination method) with the individual method is shown in Figure 8.

Our previous work [22] showed that changes in β have little effect on speech recognition performance. The spectral floor parameter influences the spectral distortion caused by the algorithm. We also conducted experiments with different spectral floor parameters for speaker recognition. The experimental results are shown in Table 8. β is the spectral floor parameter for avoiding negative or underflow values. When β is too small ($\beta = 0.05$), the dereverberation distortion is too large, worsening the results. However, if β is too large, as for $\beta = 0.25$, a lot of reverberation cannot be suppressed, so the improvement is not sufficient. Thus, we empirically set β to 0.15, which is same as for speech recognition. β is more sensitive for speaker recognition than for speech recognition.

5.2.2 Experimental results of real reverberant speech

We have verified our proposed method in a real reverberant environment. We implemented this method in a real environment using the optimal compensation parameters estimated in an artificial environment ($\alpha = n = 0.5$). The results from the real environment (Table 9) exhibited the same tendency as those in the artificial environment.

Our proposed method (*method 5*) achieved a relative error reduction of 68.3% compared with the baseline (*method 1*), and a reduction of 55.6% compared with reverberant speech models (*method 2*). For the sake of comparison, we conducted the same experiments with two other blind reverberation compensation strategies, namely LTLSS (*method 3*) and an MSLP-GSS-based method (*method 4*). The proposed method gives an error reduction rate of 35.8% compared with LTLSS and 24.7% compared with MSLP-GSS.

6 Conclusions

Previously, Wang et al. proposed a blind dereverberation method based on GSS that employed MCLMS for hands-free speech recognition [22]. In this study, we applied this method to hands-free speaker identification. However, in the speaker identification field, the method proposed in [22] performed worse than the baseline method. This is the opposite result to that for speech recognition. We addressed this problem by training speaker models using dereverberant speech, which was obtained by suppressing reverberation from arbitrary artificial reverberant speech. The reverberant speech for test data was also compensated using MCLMS-GSS-based dereverberation. By combining various compensation parameter sets for GSS and efficiently calculating the speaker likelihoods, a more robust result was obtained without parameter tuning. Based on a dereverberant speech models, the proposed method achieved a recognition rate of 93.6%, which compares well with conventional CMN with beamforming using clean speech models (49.0%), and reverberant speech models (88.4%). In addition, the method introduced in this paper does not increase the computational cost over that of previous methods. Furthermore, we implemented this method in a real environment with optimal compensation parameters estimated from an artificial environment. The proposed technique achieves a recognition rate of 87.8%, compared with 72.5% using a reverberant speech model. We also compared our proposed method with other dereverberation methods based on MSLP-GSS, both in artificial and real environments, under the same conditions of the SS method. The proposed method achieved a recognition rate of 91.7%, compared with 90.6% using MSLP-GSS, in an artificial environment, and 87.8% compared with 83.8% in a real environment.

Endnotes

[a] Delay-and-sum beamforming reduces the directivity of each microphone channel, especially when using many microphones that are far away from each other (as in the test condition). In our previous work [22], beamforming was shown to produce better results. The time delay information was calculated according to each speech recording.

[b] Details of the experimental setup are described in section 5.

[c] For real reverberant speech, the processing step is the same as for artificial reverberant speech.

[d] For example, to estimate the clean spectrum of the 2ith window W_{2i}, the estimated clean spectra of the 2(i−1)th window $W_{2(i-1)}$, the 2(i−2)th window $W_{2(i-2)}$ were used.

[e] In this study, the values of I, N, and K, in Equation 21 were set to 10, 5, and 260. γ was 92, i.e., the computational time for the likelihood calculation of K speaker models was 92 times that for feature extraction conducted on a 2.0-GHz Intel(R) Xeon(R) Server running Linux with 12-GB main memory.

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

Acknowledgements
This work was partially supported by a research grant from the Tateisi Science and Technology Foundation.

Author details

[1]Nagaoka University of Technology, Nagaoka 940-2188, Japan. [2]Shizuoka University, Hamamatsu 432-8561, Japan.

Received: 4 July 2013 Accepted: 21 December 2013
Published: 15 April 2014

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, Chiba, 26-30 September, Curran Associates, Inc., Red Hook, NY, 2010), pp. 570–573
3. C Raut, T Nishimoto, S Sagayama, Adaptation for long convolutional distortion by maximum likelihood based state filtering approach, in *Proceedings of the 2006 ICASSP* Toulouse, France, 14-19 May 2006 vol. 1 (IEEE, Piscataway, 2006), pp. 1133–1136
4. T Yoshioka, A Sehr, M Delcroix, K Kinoshita, R Maas, T Nakatani, W Kellermann, Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. **29**(6), 114–126 (2012)
5. TB Hughes, HS Kim, JH DiBiase, HF Silverman, Performance of an an HMM speech recognizer using a real-time tracking microphone array as input. IEEE Trans. Speech Audio Process. **7**(3), 346–349 (1999)
6. S Furui, Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. **29**(2), 254–272 (1981)
7. F Liu, R Stern, X Huang, A Acero, Efficient cepstral normalization for robust speech recognition, in *Proceedings of the workshop on Human Language Technology*. Princeton, 69–74 (Association for Computational Linguistics, Stroudsburg, 1993)
8. K Lebart, J Boucher, P Denbigh, A new method based on spectral subtraction for speech dereverberation. Acta Acustica. **87**, 359–366 (2001)
9. D Gelbart, N Morgan, Double the trouble: handling noise and reverberation in far-field automatic speech recognition, in *INTERSPEECH 2002* Denver, 16-20 September, 2002, pp. 968–971
10. D Gelbart, N Morgan, Evaluating long-term spectral subtraction for reverberant ASR, in *ASRU 2001* Madonna di Campiglio, Italy, 9-13 December 2001
11. M Wu, D Wang, A two-stage algorithm for one-microphone reverberant speech enhancement. IEEE Trans. ASLP. **14**(3), 774–784 (2006)
12. EA Habets, Multi-channel speech dereverberation based on a statistical model of late reverberation, in *Proceedings of IEEE ICASSP* Philadelphia, 18-23 March vol. 4, (IEEE, Piscataway, 2005), pp. 173–176
13. SO Sadjadi, JHL Hasnen, Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions, in *Proceedings of IEEE ICASSP* (Prague, Czech Republic, 22-27 May 2011), pp. 5448–5451
14. S Gannot, M Moonen, Subspace methods for multimicrophone speech dereverberation. EURASIP J. Appl. Signal Processv. **2003**(1), 1074–1090 (2003)
15. K Kinoshita, M Delcroix, T Nakatani, M Miyoshi, Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation, in *Proceedings of IEEE ICASSP 2006* Toulouse, France, 14-19 May 2006, pp. 817–820
16. S Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics Speech Signal Process. **27**(2), 113–120 (1979)
17. M Delcroix, T Hikichi, M Miyoshi, Precise dereverberation using multi-channel linear prediction. IEEE Trans. ASLP. **15**(2), 430–440 (2007)
18. Q Jin, T Schultz, A Waibel, Far-field speaker recognition. IEEE Trans. ASLP. **15**(7), 2023–2032 (2007)
19. S Subramaniam, AP Petropulu, C Wendt, Cepstrum-based deconvolution for speech dereverberation. IEEE Trans. Speech Audio Process. **4**(5), 392–396 (1996)
20. K Kinoshita, M Delcroix, T Nakatani, M Miyoshi, Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. IEEE Trans. Audio Speech Lang. Process. **17**(4), 534–545 (2009)
21. Q Jin, Y Pan, T Schultz, Far-field speaker recognition, in *Proceedings ICASSP 2006* Toulouse, France, 14-19 May vol. 1 IEEE, Piscataway, 2006), pp. 937–940
22. L Wang, K Odani, A Kai, Dereverberation and denoising based on generalized spectral subtraction by nutil-channel LMS algorithm using a small-scale microphone array. Eurasip J. Adv. Signal Process. **2012**(12) (2012)
23. 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)
24. T Inoue, H Saruwatari, Y Takahashi, K Shikano, K Kondo, Theoretical analysis of musical noise in generalized spectral subtraction based on higher-order statistics. IEEE Trans. Audio Speech Lang. Process. **19**(6), 1770–1779 (2011)
25. L Wang, S Nakagawa, N Kitaoka, Blind dereverberation based on CMN and spectral subtraction by multi-channel LMS algorithm, in *Proceedings of InterSpeech 2008* (Brisbane, 22-26, September 2008), pp. 1032–1035
26. DA Reynolds, Speaker identification and verification using Gaussian mixture speaker models. Speech Commun. **17**, 91–108 (1995)
27. DA Reynolds, TF Quatieri, R Dunn, Speaker verification using adapted Gaussian mixture models. Dig. Signal Process. **10**(1-3), 19–41 (2000)
28. L Wang, N Kitaoka, S Nakagawa, Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM. Speech Commun. **49**(6), 501–513 (2007)
29. K Farrell, R Mammone, K Assaleh, Speaker recognition using neural networks and conventional classifiers. IEEE Trans. on Speech Audio Process. **2**(1), 194–205 (1994)
30. W Campbell, J Campbell, D Reynolds, E Singer, P Torres-Carrasquillo, Support vector machines for speaker and language recognition. Comput. Speech Lang. **20**(2–3), 210–229 (2006)
31. P Kenny, P Ouellet, N Dehak, V Gupta, P Dumouchel, A study of inter-speaker variability in speaker verification. IEEE Trans. Audio Speech Lang. Process. **15**(7), 980–988 (2008)
32. N Dehak, P Kenny, R Dehak, P Dumouchel, P Ouellet, Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
33. B Kingsbury, N Morgan, Recognizing reverberant speech with RASTA-PLP, in *Proceedings of IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)* (Munich, 21-24 April vol.2 IEEE, Piscataway, 1997), pp. 1259–1262
34. AC Surendran, JL Flanagan, Stable dereverberation using microphone arrays for speaker verification. J. Acoust. Soc. Am. **96**(5), 3261–3262 (1994)
35. Y Huang, J Benesty, Adaptive blind channel identification: multi-channel least mean square and Newton algorithms, in *ICASSP* Orlando, 13-17 May vol. 2 (IEEE, Piscataway, 2002). 1637–1640
36. Y Huang, J Benesty, Adaptive multichannel least mean square and Newton algorithms for blind channel identification. Signal Process. **82**, 1127–1138 (2002)
37. 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)
38. 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)
39. S Nakamura, K Hiyane, F Asano, T Nishiura, T Yamada, Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, in *Proceedings of LREC 2000* Athens, Greece, 31 May - 2 June 2000, pp. 965–968
40. T Nishiura, M Nakayama, Y Denda, N Kitaoka, K Yamamoto, T Yamada, S Tsuge, C Miyajima, M Fujimoto, T Takiguchi, S Tamura, S Kuroiwa, K Takeda, S Nakamura, Evaluation framework for distant-talking speech recognition under reverberant environments, in *Proceedings of INTERSPEECH 2008* Brisbane, Australia, 22-26 September 2008, pp. 968–971
41. K Itou, M Yamamoto, K Takeda, T Kakezawa, T Matsuoka, T Kobayashi, K Shikano, S Itahashi, JNAS, Janpanese speech corpus for large vocabulary continuous speech recognition research. J. Acoust. Soc. Jpn. (E). **20**(3), 199–206 (1999)
42. Patrick A Naylor, Nikolay D Gaubitch, Emanuël A P Habets, Signal-based performance evaluation of dereverberation algorithms. J. Electrical Comput. Eng. **2010**(5) (2010). Article ID 127513. doi:10.1155/2010/127513

doi:10.1186/1687-4722-2014-15
Cite this article as: Zhang *et al.*: **Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:15.