Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction

Keisuke Kinoshita, *Member, IEEE*, Marc Delcroix, *Member, IEEE*, Tomohiro Nakatani, *Senior Member, IEEE*, and Masato Miyoshi, *Senior Member, IEEE*

***Abstract*—A speech signal captured by a distant microphone is generally smeared by reverberation, which severely degrades automatic speech recognition (ASR) performance. One way to solve this problem is to dereverberate the observed signal prior to ASR. In this paper, a room impulse response is assumed to consist of three parts: a direct-path response, early reflections and late reverberations. Since late reverberations are known to be a major cause of ASR performance degradation, this paper focuses on dealing with the effect of late reverberations. The proposed method first estimates the late reverberations using long-term multi-step linear prediction, and then reduces the late reverberation effect by employing spectral subtraction. The algorithm provided good dereverberation with training data corresponding to the duration of one speech utterance, in our case, less than 6 s. This paper describes the proposed framework for both single-channel and multichannel scenarios. Experimental results showed substantial improvements in ASR performance with real recordings under severe reverberant conditions.**

***Index Terms*—Automatic speech recognition (ASR), dereverberation, multi-step linear prediction (MSLP), reverberation.**

I. Introduction

A speech signal captured by a distant microphone is generally smeared by reverberation, which is caused by the reflection from, for example, walls, floors, ceilings or furniture. The reverberation is known to degrade the performance of automatic speech recognition (ASR) severely. Thus, it is desirable to find a reliable way of mitigating the effect of reverberation on ASR.

A major stream of research designed to find a way to cope with the reverberation problem involves estimating inverse filters that remove the distortion caused by the impulse response using multiple microphones. One approach for constructing such inverse filters is to first estimate the room impulse responses, and then calculate their inverse based on, for example, the multiple-input/output inverse theorem (MINT) [1]. Some researchers have proposed using a subspace method for estimating the impulse responses [2], [3]. The room impulse responses are obtained from the null space of the covariance matrix of the observed signals. However, these subspace methods are highly dependent on a prior knowledge of channel orders, and are sensitive to errors in channel order estimates. Another common approach for obtaining inverse filters is to use a linear prediction (LP) algorithm, which provides a way to calculate the inverse filter directly. Unlike the subspace approaches, LP based methods are relatively robust to channel order mismatches [4]–[6]. The dereverberation methods based on inverse filtering are developed with a solid theoretical background that enables us to achieve precise dereverberation. Therefore, they are viewed as very attractive ways of solving the reverberation problem. However, these methods are known to pose a sensitivity problem in that background noise or a small change in the transfer function results in severe performance degradation [7].

In contrast to the inverse filtering methods, robust and practical approaches have been investigated to mitigate the effect of reverberation on ASR [8]–[10]. In this paper, reverberant speech is assumed to consist of a direct-path response, early reflections and late reverberations. The early reflections are defined as the reflection components that arrive after the direct-path response within a time interval of 30 ms (which corresponds to the length of the speech analysis frame used in this paper), and the late reverberations as all the latter reflections. The early reflections may not significantly degrade ASR performance if they are handled by cepstral mean subtraction (CMS) [11] or maximum-likelihood linear regression (MLLR) [12]. On the other hand, the late reverberations can be detrimental to ASR performance [13], [14]. The standard ASR techniques to compensate the convolutional distortion such as CMS do not work well for the late reverberations. In addition, it is reported that, in a severely reverberant environment where the late reverberations have a large energy, the ASR performance cannot be improved even with an acoustic model trained with a matched reverberation condition [14]. This means that the standard acoustic model cannot handle severe late reverberations, even when they know the whole reverberation characteristics in advance. One way to resolve this is to suppress the late reverberations prior to ASR process [8]–[10]. In their studies, the energy of the late reverberations was estimated using an exponential decay function and reduced using the spectral subtraction (SS) technique [15].

Manuscript received April 09, 2008; revised September 04, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tim Fingscheidt.

K. Kinoshita, T. Nakatani, and M. Miyoshi are with the NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan (e-mail: kinoshita@cslab.kecl.ntt.co.jp; nak@cslab.kecl.ntt.co.jp; miyo@cslab.kecl.ntt.co.jp).

M. Delcroix was with the NTT Communication Science Laboratories, NTT Corporation, Kyoto 619-0237, Japan. He is now with Pixela Corporation, Osaka 556-0011 Japan (e-mail: marc@cslab.kecl.ntt.co.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2009015

The remaining early reflections are handled by CMS. Such dereverberation methods appear computationally simple and relatively robust to noise. However, since reverberation cannot be well-represented solely with such a simple model, i.e., an exponential decay model, it is difficult to achieve precise dereverberation and restore the ASR performance to the level of the recognition of clean speech.

This paper proposes a novel dereverberation method that estimates the late reverberation energy based on the concept of the inverse filtering method, namely long-term multi-step linear prediction (MSLP) [16], and performs SS to remove late reverberations, as if the desired signal and the late reverberations are uncorrelated (see Appendix I for the characteristics of late reverberations). The proposed method first uses MSLP to estimate the late reverberation signal accurately in the time domain. Then, unlike the conventional inverse filtering technique, it converts the late reverberation signal into the frequency domain, and subtracts the power spectrum of the late reverberations from that of the observed signal. In other words, while general inverse filtering methods estimate and subtract the reverberation components from the observed signal *in the time domain*, the proposed method can be interpreted as performing the subtraction *in the power spectral domain*. By excluding the phase information from the dereverberation operation based on the SS framework, the proposed method might provide a degree of robustness to certain errors that conventional sensitive inverse filtering methods could not offer. The proposed method can be formulated in either a single or multichannel scenario without major modification of the algorithm. Our experimental results revealed substantial improvements in ASR performance even in a real severe reverberant environment. The algorithm could perform good dereverberation with training data corresponding to the duration of one speech utterance, in our case, less than 6 s.

The organization of this paper is as follows. Section II introduces the signal model. In Sections III and IV, we describe the proposed dereverberation framework for single channel and multichannel scenarios. In Section V, we evaluate the proposed method in a simulated reverberant environment in terms of objective quality measurement and ASR performance. In Section VI, we perform the dereverberation of real recordings. Section VII focuses on the robustness of the proposed method in a noisy reverberant environment. Section VIII summarizes our conclusions.

In this paper, the notations $(\cdot)^T$, $(\cdot)^{-1}$, $(\cdot)^+$, $\|\cdot\|$ stand for the matrix/vector transpose, the inverse, the Moore–Penrose pseudo-inverse, and the L_2-norm, respectively. $E\{\cdot\}$ represents the time average. $\mathbf{I}$ represents the identity matrix.

II. Signal Model

We consider the acoustic system shown in Fig. 1. First, let us assume that a source signal (speech signal) $s(n)$ is produced through a Pth-order FIR filter $\mathfrak{a}(z)$ from white noise $u(n)$ as

$$s(n) = \sum_{k=0}^{P} a(k)u(n-k). \tag{1}$$

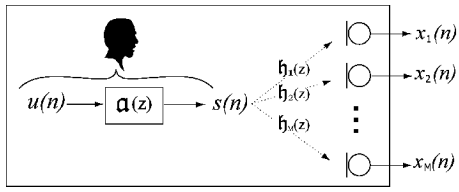

Fig. 1. Acoustic system: $u(n)$ is white noise, $\mathfrak{a}(z)$ is an FIR filter corresponding to vocal tract characteristics, $s(n)$ is a speech signal, $\mathfrak{h}_m(z)$ is the room transfer function between the speaker and the mth microphone, and $x_m(n)$ is an observed signal at the mth microphone.

where $a(n)$ is the time-domain representation of $\mathfrak{a}(z)$. Then, the speech signal recorded with a distant microphone m, $x_m(n)$, can be generally modeled as

$$x_m(n) = \sum_i h_m(i)s(n-i), \tag{2}$$

$$= \sum_l g_m(l)u(n-l). \tag{3}$$

$$g_m(l) \triangleq \sum_{k=0}^{P} h_m(l-k)a(k) \tag{4}$$

where $h_m(n)$ corresponds to the room impulse response between the source signal, and the mth microphone. $h_m(n)$ is assumed to be time invariant.

We can reformulate (3) using a matrix/vector notation as

$$\mathbf{x}_m(n) = \mathbf{G}_m\mathbf{u}(n)$$

where

$$\begin{aligned}
\mathbf{u}(n) &= [\,u(n), \quad u(n-1), \ldots, \quad u(n-T-N+1)\,]^T \\
\mathbf{x}_m(n) &= [\,x_m(n), \quad x_m(n-1), \ldots, \quad x_m(n-N)\,]^T \\
\mathbf{g}_m &= [\,g_m(0), \quad g_m(1), \quad \ldots, \quad g_m(T-1)\,] \\
\mathbf{G}_m &= \begin{bmatrix}
\mathbf{g}_m & 0 & \cdots & \cdots & 0 \\
0 & \mathbf{g}_m & \ddots & & \vdots \\
\vdots & \ddots & \mathbf{g}_m & \ddots & \vdots \\
\vdots & & \ddots & \ddots & 0 \\
0 & \cdots & \cdots & 0 & \mathbf{g}_m
\end{bmatrix}.
\end{aligned} \tag{5}$$

Here we assume $\mathbf{G}_m$ is an $(N+1)\times(T+N)$ full row rank matrix[1]. N and T indicate the dimensions of the vector $\mathbf{x}_m(n)$ and $\mathbf{g}_m$, respectively.

In this paper, a room impulse response $h_m(n)$ is assumed to consist of three parts: a direct-path response, early reflections, and late reverberations. The objective of the work described in this paper is to mitigate the effect of the late reverberations of $\mathbf{g}_m(n)$. Here let us denote the late reverberations of $\mathbf{g}_m$, $\mathbf{g}_{\mathrm{late},m}$ as

$$\mathbf{g}_{\mathrm{late},m} = [\,g_m(D), \;\; g_m(D+1), \;\; \ldots, \;\; g_m(T-1), 0, \ldots, 0\,]^T.$$

We consider that the late reverberations of $\mathbf{g}_m$ correspond to the coefficients of $\mathbf{g}_m$ after the Dth element.

[1] $\mathbf{G}_m$ is full row rank unless $\mathbf{g}_m$ is a zero matrix.

III. Single-Channel Algorithm

In this section, we introduce a dereverberation algorithm for a single-channel scenario, which represents a situation where only one observation, $x_1(n)$ in (3), is available for dereverberation.

A. Long-Term Multi-Step Linear Prediction

Here, to estimate the late reverberations, we introduce long-term multi-step LP, which was originally proposed in [16].[2]. It was first presented for the estimation of whole impulse response. In this study, we use the same method to identify only the late reverberations.

Let N be the number of filter coefficients, and D be the step-size (i.e., delay), then long-term multi-step LP can be formulated as

$$x_1(n) = \sum_{p=0}^{N} w(p)x_1(n-p-D) + e(n) \tag{6}$$

where $w(n)$ represents the prediction coefficients, and $e(n)$ is a prediction error. When D is one, the equation is equivalent to conventional LP, which is often used, for example, in speech coding and analysis [21]. The prediction coefficients can be estimated in the time domain by minimizing the mean square energy of prediction error $e(n)$. Note that these prediction coefficients are estimated based on more than at least $N + D$ samples, which amounts to several thousands in this study. In other words, the prediction coefficients are calculated using *long-term* analysis, while LPC, for example, in the speech coding field works based on short-term analysis. Using a matrix/vector notation, the obtained prediction coefficients $\mathbf{w}$ can be expressed as (see Appendix II for a detailed derivation)

$$\mathbf{w} = (\mathbf{G}_1\mathbf{G}_1^T)^{-1}\mathbf{G}_1\mathbf{g}_{\mathrm{late},1} \tag{7}$$

$$\mathbf{w} = [\, w(0), \quad w(1), \quad \ldots, \quad w(N-1)\,]^T. \tag{8}$$

Here $\mathbf{G}_1\mathbf{G}_1^T$ is a full-rank matrix because $\mathbf{G}_1$ is a full row rank matrix as mentioned above.

Now, we apply the prediction coefficients $\mathbf{w}$ to the observed signal to estimate the power of the late reverberations, as follows:

$$\begin{aligned} &E\{(\mathbf{x}_1^T\mathbf{w})^2\} \\ &= \| \mathbf{w}^T\mathbf{G}_1 E\{\mathbf{u}(n)\mathbf{u}^T(n)\}\mathbf{G}_1^T\mathbf{w} \| && (9)\\ &= \| \sigma_u^2\mathbf{w}^T\mathbf{G}_1\mathbf{G}_1^T\mathbf{w} \| . && (10)\\ &= \| \sigma_u^2\mathbf{g}_{\mathrm{late},1}^T\mathbf{G}_1^T(\mathbf{G}_1\mathbf{G}_1^T)^{-1}\mathbf{G}_1\mathbf{g}_{\mathrm{late},1} \| \\ &\le \| \sigma_u^2\mathbf{g}_{\mathrm{late},1}^T\| \cdot \|\mathbf{G}_1^T(\mathbf{G}_1\mathbf{G}_1^T)^{-1}\mathbf{G}_1\| \cdot \|\mathbf{g}_{\mathrm{late},1} \| && (11)\\ &= \| \sigma_u\mathbf{g}_{\mathrm{late},1} \|^2 . && (12) \end{aligned}$$

Using the fact that the auto-correlation matrix of white noise $u(n)$ is $E\{\mathbf{u}(n)\mathbf{u}^T(n)\} = \sigma_u^2\mathbf{I}$, where σ_u^2 is a scalar indicating the variance of $u(n)$, we can derive (10). Using the Cauchy–Schwartz inequality, we can obtain relation (11). Finally, relation (12) was obtained by using the fact that $\| \mathbf{G}_1^T(\mathbf{G}_1\mathbf{G}_1^T)^{-1}\mathbf{G}_1 \|$ is the norm of a projection matrix, which is equal to 1 [22]. Equation (12) indicates that the late reverberation components can never be overestimated in a long-term analysis sense.

Now, let us denote z-domain representation of $g_m(n)$ and $h_m(n)$ as $\mathfrak{g}_m(z)$ and $\mathfrak{h}_m(z)$. Then, as mentioned in (6) to (8), the long-term multi-step LP tries to skip the first D terms of transfer function $\mathfrak{g}_1(z)$ and estimate the remaining terms of $\mathfrak{g}_1(z)$ whose orders are higher than D. Note that $\mathfrak{g}_1(z)$ is the product of speech production system $\mathfrak{a}(z)$ and room transfer function $\mathfrak{h}_1(z)$ as in (4). Therefore, the late reverberation energy calculated as in (12) may include not only the contribution of the late reverberations of $\mathfrak{h}_1(z)$ but also the bias caused by $\mathfrak{a}(z)$. In order to reduce this bias, we suggest employing a preprocessing technique for long-term multi-step LP, known as the pre-whitening technique, which appears to be effective in reducing the short-term correlation of a speech signal produced through $\mathfrak{a}(z)$. In this paper, this pre-whitening was done by using small order LP ($\simeq$ 20 taps), which can be estimated as shown in Appendix III. Care has to be taken to choose the LP order for long-term multi-step LP and pre-whitening. The long-term multi-step LP tries to model the late reverberations of $\mathfrak{h}_1(z)$; thus, the order N has to be very high. In contrast, the LP order used for pre-whitening should be small, since the aim of this processing is only to suppress the short-term correlation caused by speech production system $\mathfrak{a}(z)$.

B. Spectral Subtraction

Here we propose the use of SS to suppress the late reverberations. That is, we first divide the observed signal and the estimated late reverberations into short frames, apply short-term Fourier transform (STFT) to calculate the power spectrum, and then subtract the power spectrum of the estimated late reverberations from that of the observed signal. Although, in the previous section, we showed that the power of the predicted late reverberations can never be overestimated compared with that of true late reverberations in the long-term analysis sense, some degree of overestimation may occur in (short-term) local time region.

In summary, an exact subtraction rule can be formulated as shown below, by denoting the STFT of a short segment of the observed signal at the mth microphone as $X_m(k\lambda,\omega)$ and that of the estimated late reverberations as $R_m(k\lambda,\omega)$, where λ is the frame length and k is an integer

$$|\hat{S}_m(k\lambda,\omega)| = \begin{cases} \sqrt{|X_m(k\lambda,\omega)|^2 - |R_m(k\lambda,\omega)|^2}, \\ \qquad (\text{if} |X_m(k\lambda,\omega)|^2 - |R_m(k\lambda,\omega)|^2 \ge 0) \\ 0, \quad (\text{otherwise}) \end{cases}$$

where $\hat{S}_m(k\lambda,\omega)$ denotes the STFT of the dereverberated signal. To synthesize a time-domain dereverberated signal, we simply apply the phase of the observed signal $\angle X_m(kM,\omega)$ as

$$\hat{S}_m(k\lambda,\omega) = |\hat{S}_m(k\lambda,\omega)|e^{j\angle X(k\lambda,\omega)}.$$

[2]There are several speech dereverberation methods that also use LP [17]–[20] Note that, in their studies, LP was mainly used to model speech components, thus the LP order is relatively small ($\simeq$ 20). In contrast, here we wish to model reverberation with long-term multi-step LP; thus, the order is much higher (i.e., several thousands).

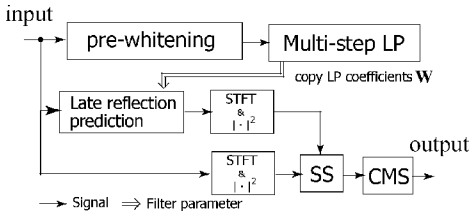

Fig. 2. Schematic diagram of proposed method for single-channel scenario.

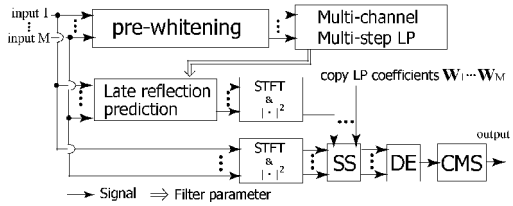

Fig. 3. Schematic diagram of multichannel implementation.

C. Schematic Processing Diagram of Single-Channel Algorithm

Fig. 2 is a schematic diagram of the proposed method for a single-channel scenario mentioned above. First the observed signal is prewhitened with small order LP, and processed with the long-term multi-step LP. The long-term multi-step LP is used to obtain the coefficients $\mathbf{w}$ that best predict the late reverberations. Then, by convoluting (or filtering) the observed signal with the prediction coefficients as $\mathbf{x}_1^T\mathbf{w}$, we estimate the late reverberations. After applying a STFT to the observed signal and predicted late reverberations, we perform SS in the the spectral domain to remove the effect of the late reverberations from the observed signal (shown as "SS" in Fig. 2) [15]. Finally, to remove the remaining early reflections for the ASR system, we apply CMS to the processed signal.

IV. Multichannel Algorithm

In this section, we extend the proposed algorithm to the multichannel scenario. By employing the multichannel long-term multi-step LP [16], the two sides of (12) become equal [1], [23]; thus, we expect to estimate the late reverberations more accurately.

A. Multichannel Long-Term Multi-Step Linear Prediction

Here, we introduce multichannel long-term multi-step LP to estimate late reverberations based on multiple observed signals, Let L be the number of filter coefficients for each channel, D be the step-size (i.e., delay), and M be the number of microphones, then the multichannel long-term multi-step LP is formulated as follows:

$$x_i(n) = \sum_{m=1}^{M}\sum_{p=0}^{L} w_{m,i}(p)x_m(n-p-D) + e_i(n), \quad (i = 1,2\cdots M) \tag{13}$$

where $x_m(n)$ corresponds to the observed signal at the mth microphone, and $w_{m,i}(n)$ to the prediction coefficients at the mth microphone when the prediction target is the observed signal at the ith microphone $x_i(n)$. The multichannel long-term multi-step LP calculates the late reverberations within $x_i(n)$. The prediction coefficients $w_{m,i}(n)$ can be estimated by minimizing the mean square energy of the prediction error $e_i(n)$ (see Appendix IV for a detailed derivation). Using a matrix/vector notation, the obtained prediction coefficients $\mathbf{w}_i$ can be written in a similar manner to the single channel algorithm as:

$$\mathbf{w}_i = (\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{G}\mathbf{g}_{\mathrm{late},i}, = (\mathbf{G}^T)^{+}\mathbf{g}_{\mathrm{late},i} \tag{14}$$

where

$$\mathbf{w}_{m,i} = [\, w_{m,i}(0) \quad ,\ldots, \quad w_{m,i}(L-1)]^T,$$
$$\mathbf{w}_i = [\, \mathbf{w}_{1,i}{}^T, \mathbf{w}_{2,i}{}^T \quad ,\ldots, \mathbf{w}_{M,i}{}^T\,]^T,$$
$$\mathbf{G} = [\, \mathbf{G}_1^T, \quad \mathbf{G}_2^T \quad ,\ldots, \quad \mathbf{G}_m^T\,]^T.$$

Now, let us apply the prediction coefficients $\mathbf{w}_i$ to the observed signal to estimate the late reverberations. Here, we define the observed signal $\mathbf{x}(n)$ as

$$\mathbf{x}(n) = [\, \mathbf{x}_1^T(n), \quad \mathbf{x}_2^T(n) \quad ,\ldots, \quad \mathbf{x}_M^T(n)\,]^T.$$

Then, the estimated late reverberations can be expressed as follows:[3]

$$\begin{aligned}\mathbf{x}^T(n)\mathbf{w}_i &= \mathbf{u}^T(n)\mathbf{G}^T\mathbf{w}_i,\\ &= \mathbf{u}^T(n)\mathbf{G}^T(\mathbf{G}^T)^{+}\mathbf{g}_{\mathrm{late},i},\\ &\simeq \mathbf{u}^T(n)\mathbf{g}_{\mathrm{late},i}.\end{aligned} \tag{15}$$

Equation (15) simply indicates that the late reverberations can be more accurately estimated. In other words, now with multichannel long-term multi-step LP, the two sides of (12) become the same.

B. Schematic Processing Diagram

Fig. 3 shows an algorithm based on the multichannel long-term multi-step LP. There are two major modifications compared with the single-channel algorithm. First, in the multichannel scenario, we perform long-term multi-step LP based on signals captured by multiple microphones. Second, to enhance the direct-path response in the processed speech, we adjust the delays and calculate the sum of the signals from all the channels. The process is denoted as "Direct-path Enhancement (DE)" in the figure.

First, pre-whitening is applied to each of the observed signals. Next, using multichannel long-term multi-step LP, we estimate the late reverberations at the ith microphone. Based on the STFT of the estimated late reverberations and that of the observed signals, we calculate the dereverberated signal at the ith microphone. We repeat this procedure for all i ($i = 1,2\cdots M$) to obtain the dereverberated speech for all the microphones. Then, we adjust the delays among the output signals and calculate their sum to obtain the resultant signal. The delays were estimated with the Generalized Cross-Correlation (GCC) method [24]. Finally, to remove the remaining early reflections, we apply CMS to the processed signal.

[3]For (15) to be strictly equal, $\mathbf{H}$, which is the Sylvester matrix of $h_m(n)$, similar to $\mathbf{G}$, has to be a full column rank matrix.

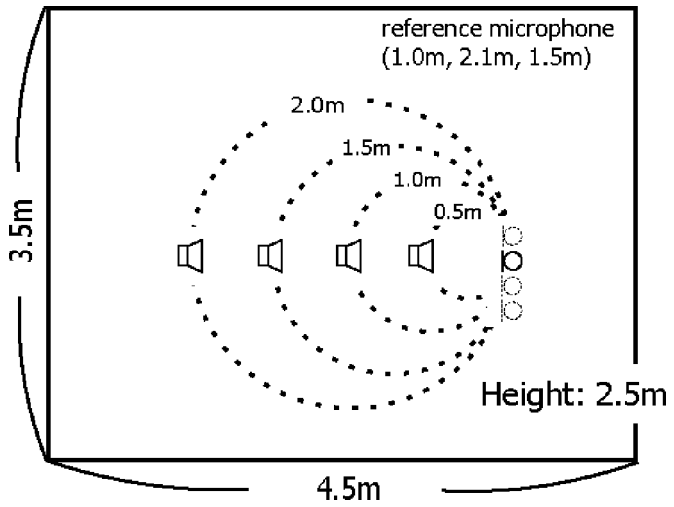

Fig. 4. Experimental setup: the reflection coefficients of the walls are [0.93 0.93 0.94 0.94 0.15 0.15].

TABLE I
EXPERIMENTAL CONDITIONS FOR ASR

Sampling rate	12 kHz (16-bit quantization)
Feature vector (39 dimensions)	12 - order MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Energy + Δ Energy + $\Delta\Delta$ Energy
Window	Hamming
Frame size/shift	30/10 ms
Number of states	3 (Left to right)
Number of phoneme categories	43
Number of clustered states	3000

TABLE II
TRAINING AND TEST DATA FOR ACOUSTIC MODEL AND LANGUAGE MODEL FOR JNAS

Training data for female speakers	JNAS: 20,103 utterances, 33 hours (131 females)
Training data for male speakers	JNAS: 20,093 utterances, 34 hours (122 males)
Test data for female speakers	JNAS: 100 utterances, 1,578 words (22 females)
Test data for male speakers	JNAS: 100 utterances, 1,578 words (22 males)
Language model	Standard trigram (10 years of Japanese newspapers)
Vocabulary size	20,000
Perplexity	94.8 (OOV rate:2.1 %)

V. EXPERIMENT IN SIMULATED REVERBERANT ENVIRONMENT

In this section, we evaluate the effectiveness of the proposed methods in a simulated reverberant environment, where our *noise-free* assumption holds.

A. Experimental Conditions

1) Reverberation Condition: Fig. 4 summarizes the acoustic environment for the experiment. The single-channel processing employed the microphone shown with the solid line, while the four-channel processing employed three extra microphones indicated with dotted lines. Each microphone was equally spaced at a distance of 0.2 m. Impulse responses were simulated with the image method [25], for four different speaker positions, with distances of 0.5, 1.0, 1.5, and 2.0 m between the reference microphone and the speaker. The RT_{60} reverberation time of the simulated acoustic environment was about 0.65 s[4]. The impulse response was 9600 taps corresponding to a duration of 0.8 s, with a sampling frequency of 12 kHz.

2) ASR Condition: The Japanese Newspaper Article Sentences (JNAS) corpus was used to investigate the effectiveness of the proposed method as a preprocessing algorithm for ASR. The ASR performance was evaluated in terms of word error rate (WER) averaged over genders and speakers. In the acoustic model, we used the following parameters : 12 order MFCCs + energy, their Δ and $\Delta\Delta$, three state triphone HMMs, and 16 mixture Gaussian distributions. The acoustic model settings are summarized in Table I. The total number of clustered states was set at 3000 using a decision-tree based context clustering technique [27]. The model was trained on clean speech processed with CMS. The language model was a standard trigram trained on Japanese newspaper articles written over a ten-year period. The training and test sets for the recognition task are summarized in Table II. The duration of the test data ranged from 2 to 16 s, and the average value was about 6 s.

[4]In [26], we carried out experiments with RT_{60} values of 0 to 0.5 s.

3) Parameters for Dereverberation: The filter length for single-channel algorithm N, that for multichannel algorithm L, and the step-size D in (6) and (13), were 3000, 750, and 360, respectively. It should be noted that, when dealing with longer reverberations, in theory we simply have to use a longer filter. Here, D is set at the length of the analysis frame used for CMS to deal with all the reverberation components that CMS cannot handle. For the pre-whitening, we used 20th-order LP, which we calculated similarly to the approach described in [20] (see Appendix III for details). In our experiment, the coefficients of the pre-whitening filter were fixed for an entire utterance. Although we determined these orders experimentally, according to the preliminary experiment, we confirmed that similar performance could be obtained for different filter lengths N given a range of 1000 taps. No special parameters were used for spectral subtraction. These parameters are common to all the experiments reported in this paper.

The dereverberation was performed utterance by utterance. The estimation of the LP coefficients starts only after all samples corresponding to the current utterance become available. This means that the length of the training data used to estimate the LP coefficients is equivalent to the duration of each input utterance. We have confirmed experimentally that, if we can use the data of more than about 2 s of data, we can obtain sufficiently converged LP coefficients, and the algorithm performance become relatively stable. We employed the Levinson–Durbin algorithm for single-channel long-term multi-step LP [21], and the class of Schur's algorithm for multichannel long-term multi-step LP [21], [28]–[30] to calculate the prediction coefficients efficiently. These fast algorithms enable us to run the whole process at a real-time factor of less than 1, for example, on the Intel Pentium IV 3.4-GHz processor used in our experiments.

When we compare the length of the simulated impulse responses and the filter length for MSLP, we find that the current filter length is not sufficiently long to estimate all the late reverberations, and the analysis of the proposed dereverberation method presented in Sections III and IV does not hold precisely.

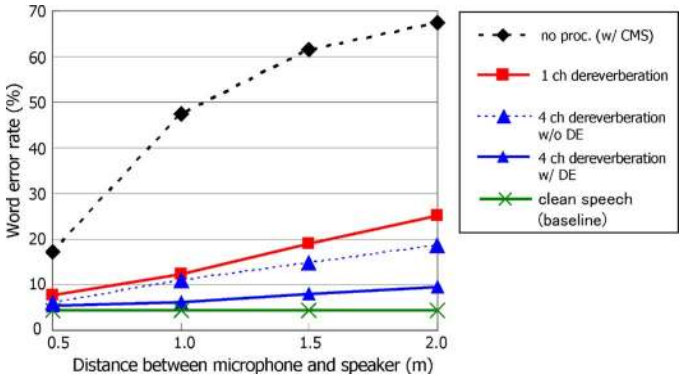

Fig. 5. Recognition experiment in a simulated reverberant environment: Recognition performance as a function of the distance between microphone and speaker.

However, we chose this filter length to allow us to execute the whole process in a realistic computational time.

B. Dereverberation Effect on ASR

Fig. 5 shows the WER as a function of the distance between the microphone and the speaker. "No proc." corresponds to the WER of the reverberant speech processed with CMS, "1 ch dereverberation" to that of speech dereverberated with the single channel algorithm, "4 ch dereverberation w/ DE" to that of speech dereverberated with the four channel algorithm with the DE process (as shown in Fig. 3). "4 ch dereverberation w/o DE" is the signal of one representative channel that was captured immediately before being passed to the DE process in the process of the four channel algorithm. This example is provided to show the improvement that we can gain by extending single channel long-term multi-step LP to multichannel form. "Clean speech (baseline)" is the lowest possible WER, i.e., 4.4%, that can be realized with this ASR system based on this corpus.

As seen from the figure, if the reverberant speech undergoes no preprocessing, the WER increases greatly as the distance increases. With the proposed method, we achieved a substantial reduction in the WER with both the single channel and four channel algorithms for all reverberant conditions. The improvement obtained by using four channels rather than a single-channel becomes more obvious, particularly as the distance between the speaker and the microphone increases.

C. Spectrogram Improvement

Fig. 6 shows a spectrogram of clean speech processed with CMS, reverberant speech at a distance of 1.5 m, speech dereverberated by the single-channel algorithm, speech dereverberated by the four-channel algorithm without the DE process, and speech dereverberated by the four-channel algorithm with the DE process. We can clearly see the effect of the proposed method in both the single-channel and four-channel cases. Although we can observe some differences between the levels of performance provided with the single-channel and four-channel algorithms, no significant improvement can be seen in spectrograms. Although (12) implicitly shows that the single-channel algorithm may greatly underestimate the power of late reverberations, this experimental result supports the idea that the

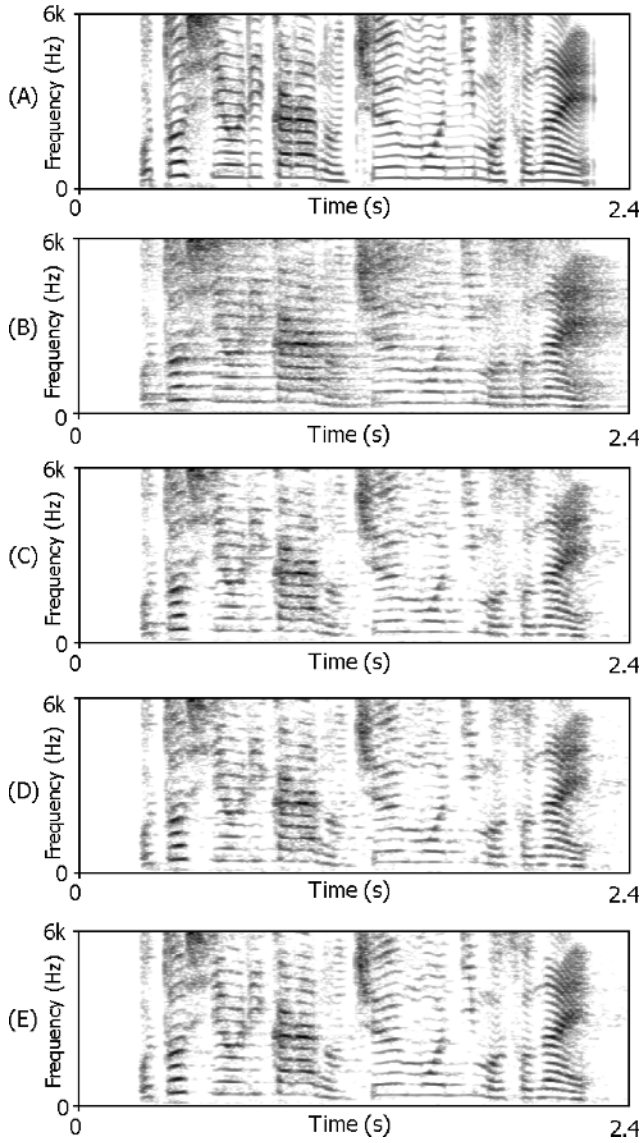

Fig. 6. Spectrograms in a simulated reverberant environment when the distance between the microphones was set at 1.5 m: (A) clean speech, (B) reverberant speech, (C) speech dereverberated by the single-channel algorithm, (D) speech dereverberated by the four-channel algorithm without DE, and (E) speech dereverberated by the four-channel algorithm with DE.

algorithm successfully generates a reasonable estimate of late reverberations. Note that, since no over-subtraction factor is used in the present work, if the power of late reverberations is greatly underestimated, a spectrogram should show some evidence of the remaining late reverberations.

D. Evaluation With LPC Cepstrum Distance

Here we use the average LPC cepstrum distance [31] to evaluate the precision of the dereverberation with an objective measurement. Fig. 7 shows the average LPC cepstrum distance between clean speech processed with CMS and target speech. To calculate the LPC cepstrum distance, we excluded the silence found at the beginning and end of the utterance files. The legends represent the same type of speech signal as those in Fig. 5. Here again, the difference in performance between single-channel and four-channel processing becomes more

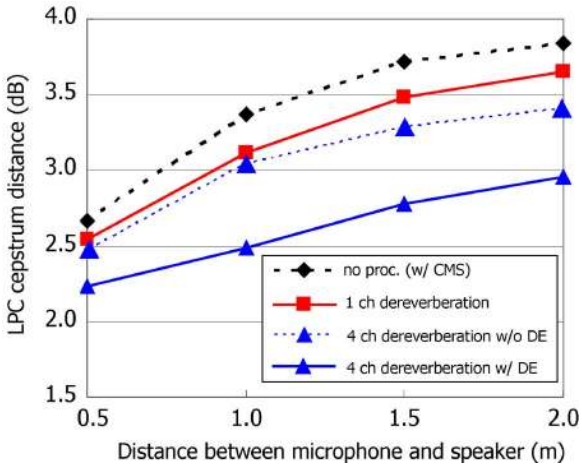

Fig. 7. LPC cepstrum distance in simulated reverberant environment as a function of the distance between the microphone and the speaker.

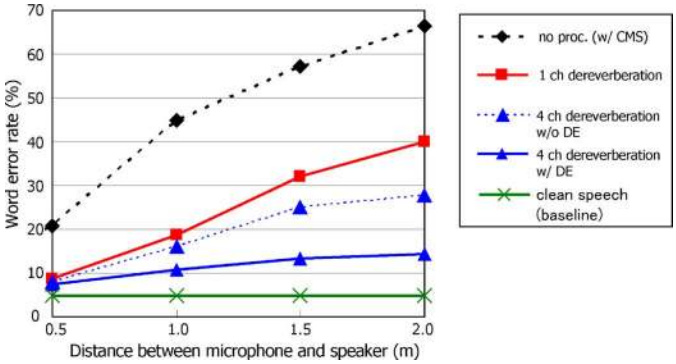

Fig. 8. Recognition experiment in real reverberant environment: Recognition performance as a function of the distance between the microphone and the speaker.

noticeable as the distance increases, as previously noticed in Fig. 5.

VI. Experiment in Real Reverberant Environment

In this section, we carried out experiments with speech recorded in a real reverberant room to show the applicability of the proposed method.

A. Experimental Condition

The recordings were made in a reverberant chamber with the same dimensions as the simulated room described in Section IV. The location of the microphones and loudspeaker also follows the simulation setup depicted in Fig. 4. For each gender, 100 Japanese sentences taken from the JNAS database were played through a BOSE 101VM loudspeaker, and recorded with SONY ECM-77B omnidirectional microphones. The positions of the loudspeaker and the microphones were fixed throughout the recordings. The signal-to-noise ratios (SNRs) of the recordings were about 15 to 20 dB, and the RT_{60} reverberation time was about 0.5 s. The D_{50} values are approximately the same as those of simulated impulse responses [32]. We applied high-pass filtering to the recordings before the dereverberation process to suppress the unwanted background noise, which was mainly concentrated below 200 Hz. After the high-pass filtering, the SNRs were about 30 dB. As a control, we also recorded the same utterances in a nonreverberant chamber with a close microphone using the same experimental equipment.

B. Dereverberation Effect on ASR

We also carried out ASR experiments with real recordings. The acoustic and language models were the same as in Section V. The training and test sets for this recognition task were the same as for the previous experiment and are summarized in Table II.

Fig. 8 shows the WER of the real recordings as a function of the distance between the microphone and the speaker. The legends represent the same type of processing as those in Fig. 5. In this experiment, the baseline performance is 4.9%, which is the WER obtained with recordings made in a nonreverberant chamber.

The improvement in WER is sufficiently noticeable under all reverberant conditions, and the global tendency is similar to the simulation. The results indicate that the proposed framework also works well even with speech recorded in a severely reverberant environment.

C. Spectrogram Improvement

In this experiment, to move one step nearer a real scenario, we attempted the dereverberation of actual human utterances (rather than those from loudspeaker). In this case, the source position might be constantly fluctuating owing to head movement, despite the speaker being asked to stand still during the recordings at the same position as the loudspeaker in Fig. 4.

Fig. 9 shows spectrograms of recorded reverberant speech uttered by a male speaker, speech dereverberated with the single-channel algorithm, speech dereverberated by the four-channel algorithm without the DE process, and speech dereverberated by the four-channel algorithm with the DE process. Here, we again see the substantial reduction in reverberation in both the single- and four-channel cases.

VII. Robustness of Proposed Dereverberation Method to Diffusive Noise

In this section, we evaluate our proposed method under noisy reverberant conditions to confirm its robustness. The evaluations are undertaken using spectrograms and LPC cepstrum distance. To perform an ASR test in a noisy environment, the method should be combined with noise adaptation techniques such as spectral subtraction [15] and parallel model combination [33], [34]. Since we would like to focus primarily on the reverberation problem in this paper, we do not include the issue of combining the proposed method with other noise adaptation techniques. Please refer to [35] for an evaluation of the proposed dereverberation method combined with SS [15] in a noisy reverberant environment.

A. Experimental Condition

The reverberation conditions are the same as those described in Section V. To simulate an environment with diffusive noise,

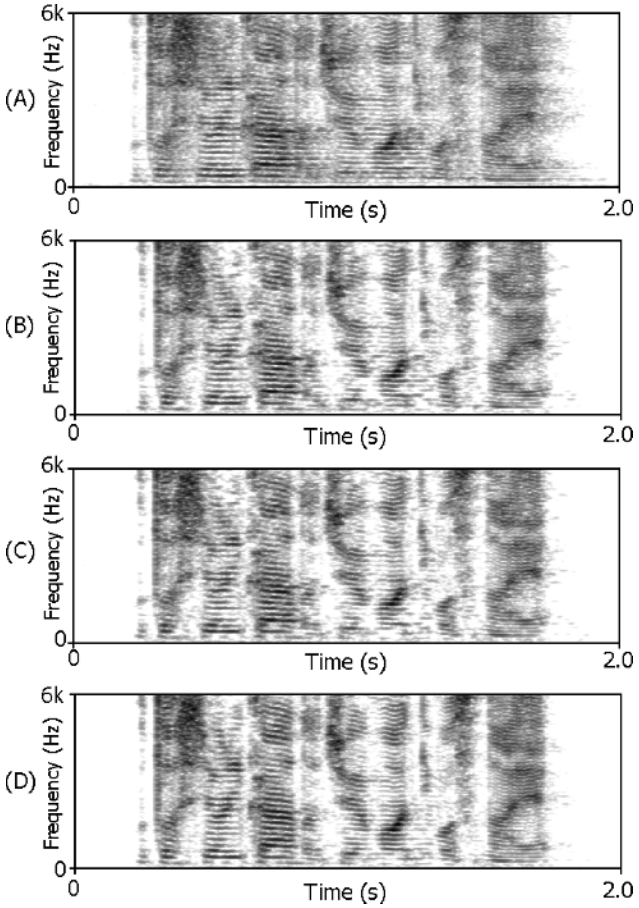

Fig. 9. Spectrograms obtained in a real reverberant environment when the distance between the microphones and speaker was set at 1.5 m: (A) recorded reverberant speech, (B) speech processed with the single-channel algorithm (C) speech dereverberated by the four-channel algorithm without the DE process, and (D) speech dereverberated by the four-channel algorithm with the DE process.

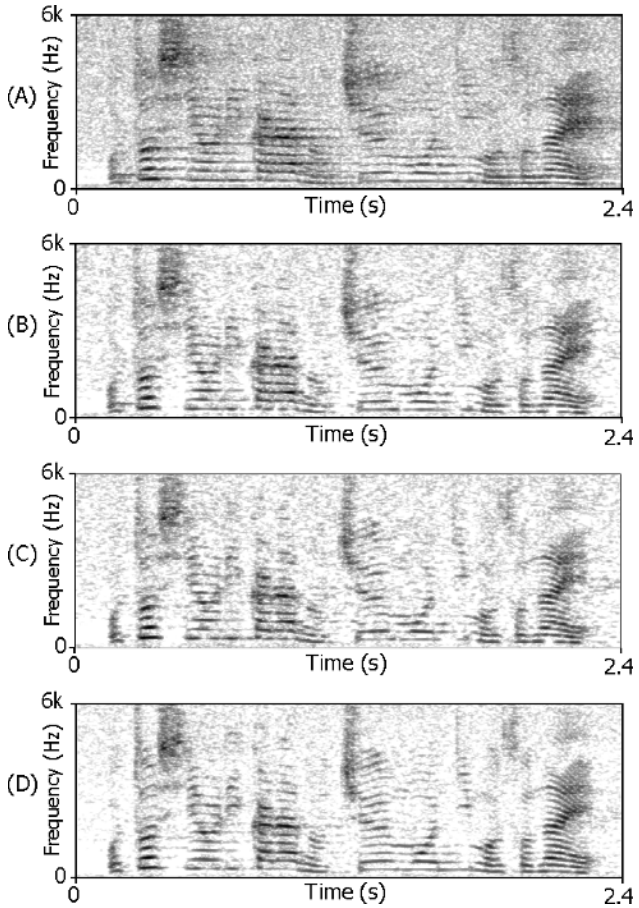

Fig. 10. Spectrograms in a noisy reverberant environment, when the distance between the microphones and speaker was set at 1.5 m, and the SNR was 20 dB: (A) noisy reverberant speech, (B) speech dereverberated by the single-channel algorithm, (C) speech dereverberated by the four-channel algorithm without DE, and (D) speech dereverberated by the four-channel algorithm with DE.

white noise was artificially generated and added to reverberant speech with SNRs of 0, 10, 20, 30, and 40 dB.

B. Spectrogram Improvement

Fig. 10 shows spectrograms of the observed noisy reverberant speech, speech dereverberated by the single-channel algorithm, speech dereverberated by the four-channel algorithm without the DE process, and with the DE process with a 20-dB SNR. Here, the distance between the speaker and the microphones was set at 1.5 m. From the spectrograms, we could see that both single-channel and four-channel dereverberation works fairly well even in a noisy environment. It may be interesting to note that, although the algorithm does not explicitly perform denoising, some denoising effect can be seen especially in Fig. 10 (D). This is probably due to the DE processing employed with the four-channel algorithm.

C. Evaluation With LPC Cepstrum Distance

Here, to evaluate the dereverberation precision in a noisy environment, we calculated the LPC cepstrum distance between clean speech processed with CMS and the target speech. In this case, the dereverberated speech was generated by estimating the LP coefficients in a noisy environment, and then processing the noiseless reverberant speech with the coefficients. By doing this, the dereverberation performance could be evaluated without taking account of the spectral distortion caused by the background noise. The results are summarized in Fig. 11, where the legends represent the same type of processing as those in Fig. 5. Note that, the 40-dB SNR case shown in Fig. 11 approximately coincide with Fig. 7, which shows the case of $+\infty$ SNR. The proposed method appears to provide stable performance for SNRs above 20 dB. Even though the accuracy decreases for SNRs below 20 dB, the dereverberation effect is still noticeable when using the four-channel algorithm with DE. Consequently, the proposed framework is relatively robust to background noise.

VIII. Conclusion

A speech signal captured by a distant microphone is generally smeared by reverberation, which severely degrades the ASR performance. In this paper, we proposed a novel dereverberation method that combines the concept of inverse filtering and well-known spectral subtraction. The method first estimates late reverberations using long-term multi-step linear prediction, and then suppresses them with subsequent spectral subtraction.

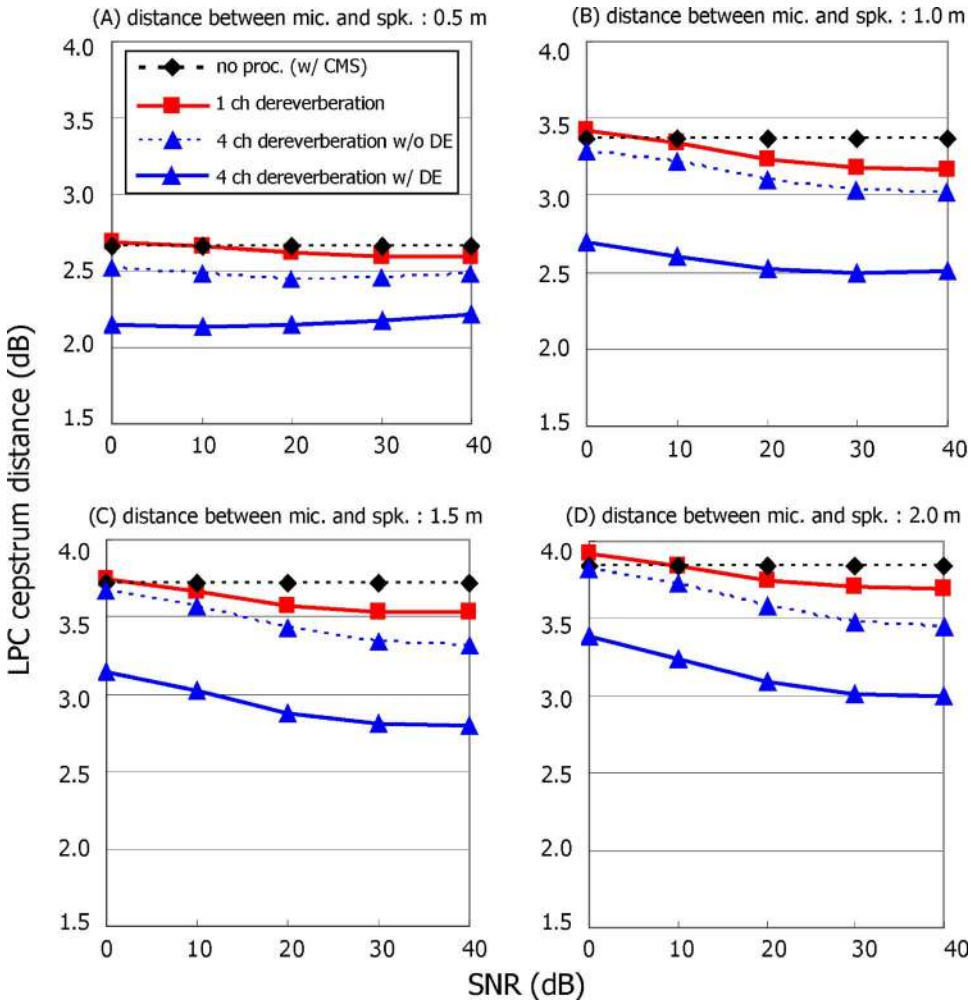

Fig. 11. LPC cepstrum distance as a function of SNR : Each panel is different as regards the distance between the microphone and the speaker. The top left and right panels and the bottom left and right panels correspond to 0.5, 1.0, 1.5, and 2.0 m, respectively.

Experimental results showed that both single and multichannel algorithms could achieve good dereverberation and could significantly improve the ASR performance even in a real severe reverberant environment. In particular, with the multichannel algorithm, the recognition performance was sufficiently close to an anechoic scenario. Since the multichannel algorithm can estimate the late reverberations more accurately compared to the single-channel one and can be advantageously combined with the postprocessing to enhance the direct-path response, it allowed us to perform more efficient dereverberation. We also discussed the robustness of the proposed method to white background noise, and confirmed that the performance was stable for SNRs above 20 dB.

In future work, we will consider the effect of background noise explicitly, and achieve not only dereverberation but also denoising.

APPENDIX I
CHARACTERISTICS OF LATE REVERBERATIONS

Here let us describe the characteristics of late reverberations and their relationship to direct-path response and early reflections.

A speech signal has a strong correlation within each local time region due to articulatory constraints, and it loses the correlation as a result of articulatory movements. Therefore, it may be possible to assume that the autocorrelation of clean speech $s(n)$, $R_{ss}(\tau) = E\{s(n)s(n-\tau)\}$, has the following property:

$$|R_{ss}(\tau)| \simeq 0 \quad \text{iff } \tau \geq \tau_s \tag{16}$$

where, with a speech signal, the value τ_s can vary approximately from 30 to 100 ms depending on the phoneme of interest.

Using τ_s and the length of the room impulse response κ, we rewrite (2) as

$$x_m(n) = \sum_{i}^{\kappa} h_m(i)s(n-i) \tag{17}$$

$$= h_m(0)s(n) + \sum_{k=1}^{\tau_s-1} h_m(k)s(n-k) + \sum_{k=\tau_s}^{\kappa} h_m(k)s(n-k). \tag{18}$$

If τ_s is equivalent to 30 ms (which corresponds to the length of the speech analysis frame in this paper), the second and third

terms of (18) exactly coincide with the definitions of the early reflections and late reverberations, respectively. If we assume the condition of (16), we can assume the late reverberations to be uncorrelated with the direct-path response, and if $\kappa \gg \tau_s$ and $\sum_{k=\tau_s}^{\kappa} h_m$ has sufficient energy, it may be possible to assume that the late reverberations and early reflections are also uncorrelated.

Appendix II
Derivation of Prediction Coefficients in Single-Channel Scenario

By minimizing the mean square energy of the prediction error $e(n)$ in (6), we could obtain the prediction coefficients. Using matrix/vector notation, the minimization of $e(n)$ leads to the following equation:

$$(E\{\mathbf{x}_1(n-D)\mathbf{x}_1^T(n-D)\})\mathbf{w} = E\{\mathbf{x}(n-D)x_1(n)\} \quad (19)$$

where

$$\mathbf{w} = [\, w(0), \quad w(1), \quad \ldots, \quad w(N-1)\,]^T.$$

Thus, the prediction coefficients can be obtained as

$$\mathbf{w} = (E\{\mathbf{x}_1(n-D)\mathbf{x}_1^T(n-D)\})^{-1}E\{\mathbf{x}(n-D)x_1(n)\}. \quad (20)$$

To understand the behavior of $\mathbf{w}$, we now expand (20). First, the term in $(\cdot)^{-1}$ can be expanded as

$$\begin{aligned} E\{\mathbf{x}_1(n-D)\mathbf{x}_1^T(n-D)\} &= \mathbf{G}_1 E\{\mathbf{u}(n-D)\mathbf{u}^T(n-D)\}\mathbf{G}_1^T, \\ &= \sigma_u^2 \mathbf{G}_1\mathbf{G}_1^T \end{aligned}$$

where the auto-correlation matrix of white noise $u(n)$ $E\{\mathbf{u}(n-D)\mathbf{u}^T(n-D)\}$ is assumed to be $\sigma_u^2\mathbf{I}$. σ_u^2 is a scalar that corresponds to the variance of $u(n)$. The second term can also be expanded as

$$\begin{aligned} E\{\mathbf{x}(n-D)x_1(n)\} &= \mathbf{G}_1 E\{\mathbf{u}(n-D)\mathbf{u}^T(n)\}\mathbf{g}_1^T, \\ &= \sigma_u^2 \mathbf{G}_1\mathbf{g}_{\mathrm{late},1}^T. \end{aligned}$$

Finally $\mathbf{w}$ can be rewritten as

$$\mathbf{w} = (\mathbf{G}_1\mathbf{G}_1^T)^{-1}\mathbf{G}_1\mathbf{g}_{\mathrm{late},1} \quad (21)$$

where

$$\mathbf{g}_{\mathrm{late},1} = [\, g_1(D), \quad g_1(D+1), \quad \ldots, \quad g_1(T-1), 0, \ldots, 0\,]^T.$$

Here, we consider that the late reverberations correspond to the coefficients of $g_1(n)$ after the Dth element, and are represented by $\mathbf{g}_{\mathrm{late},1}$.

It should be noted that (19) can be solved efficiently, for example, by the Levinson–Durbin algorithm [21].

Appendix III
Estimation of Pre-Whitening Filter

In this paper, the following qth-order prediction filter $\alpha(n)$ was used for pre-whitening to equalize $\mathfrak{a}(z)$ in (1). We first calculate the auto-correlation coefficient with the lag of c samples using the observed signal at the mth microphone as

$$r_m(c) = E[x_m(n)x_m(n+c)] \quad (c = 0, 1, 2\cdots). \quad (22)$$

Then, we take the average of $r_m(c)$ over all the channels.

$$\phi(c) = \frac{1}{M}\sum_{m=1}^{M} r_m(c). \quad (23)$$

As with standard LP [21], using $\phi(c)$, the prediction filter $w(n)$ was calculated based on the following Yule–Walker equation:

$$\begin{bmatrix} \alpha(1) \\ \alpha(2) \\ \cdot \\ \cdot \\ \alpha(q) \end{bmatrix} = \begin{bmatrix} \phi(0) & \phi(1) & \cdots & \cdots & \phi(q-1) \\ \phi(1) & \phi(0) & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \phi(1) \\ \phi(q-1) & \cdots & \cdots & \phi(1) & \phi(0) \end{bmatrix}^{-1} \times \begin{bmatrix} \phi(1) \\ \phi(2) \\ \cdot \\ \cdot \\ \phi(q) \end{bmatrix}. \quad (24)$$

Appendix IV
Derivation of Prediction Coefficients in Multichannel Scenario

By minimizing the mean square energy of the prediction error $e_i(n)$ in (13), we could obtain the prediction coefficients. The minimization of $e_i(n)$ leads to the following equation:

$$(E\{\mathbf{x}(n-D)\mathbf{x}^T(n-D)\})\mathbf{w}_i = E\{\mathbf{x}(n-D)x_i(n)\} \quad (25)$$

where

$$\begin{aligned} \mathbf{x}(n) &= [\, \mathbf{x}_1^T(n), \quad \mathbf{x}_2^T(n), \quad \ldots, \quad \mathbf{x}_M^T(n)\,]^T \\ \mathbf{w}_{m,i} &= [\, w_{m,i}(0), \quad \ldots, \quad w_{m,i}(L-1)\,]^T \\ \mathbf{w}_i &= [\, \mathbf{w}_{1,i}{}^T, \mathbf{w}_{2,i}{}^T, \quad \ldots, \mathbf{w}_{M,i}{}^T\,]^T \end{aligned}$$

Thus, $\mathbf{w}_i$ can be obtained as

$$\mathbf{w}_i = (E\{\mathbf{x}(n-D)\mathbf{x}^T(n-D)\})^{+}E\{\mathbf{x}(n-D)x_i(n)\}. \quad (26)$$

To understand the behavior of $\mathbf{w}_i$, we reformulate (26) in a similar manner to that used for a single-channel and described above. Now, $\mathbf{w}_i$ can be rewritten as

$$\begin{aligned}\mathbf{w}_i &= (\mathbf{G}\mathbf{G}^T)^{+}\mathbf{G}\mathbf{g}_{\mathrm{late},i} \\ &= (\mathbf{G}^T)^{+}\mathbf{g}_{\mathrm{late},i}\end{aligned} \tag{27}$$

where

$$\begin{aligned}\mathbf{G} &= [\,\mathbf{G}_1^T, \quad \mathbf{G}_2^T, \quad \ldots, \quad \mathbf{G}_M^T\,]^T \\ \mathbf{g}_{\mathrm{late},i} &= [\,g_i(D), \quad g_i(D+1), \quad \ldots, \quad g_i(T-1), 0, \ldots, 0\,]^T.\end{aligned}$$

Note that, (25) can be efficiently solved by, for example, the class of Schur's algorithm, which is able to determine a least square solution for general block–Toeplitz matrix equations [21], [28]–[30].

REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[2] M. I. Gurelli and C. L. Nikias, "EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 134–149, Jan. 1995.

[3] S. Gannot and M. Moonen, "Subspace methods for multi microphone speech dereverberation," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[4] J. Ayadi and D. T. M. Slock, "Multichannel estimation by blind MMSE ZF equalization," in *Proc. 2nd IEEE Workshop Signal Process. Adv. Wireless Commun.*, 1999, pp. 251–254.

[5] L. Tong and Q. Zhao, "Joint order detection and blind channel estimation by least squares smoothing," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2345–2355, Sep. 1999.

[6] G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, *Signal Processing Advances in Wireless and Mobile Communications*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[7] B. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 311–319, May 2000.

[8] K. Lebart and J. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoust.*, vol. 87, pp. 359–366, 2001.

[9] I. Tashev and D. Allred, "Reverberation reduction for improved speech recognition," in *Proc. Hands-Free Commun. Microphone Arrays*, 2005, pp. 8–9.

[10] M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, vol. 1, pp. 844–847.

[11] T. F. Quatieri, *Discrete-Time Speech Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice-Hall, 1997.

[12] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech, Lang.*, vol. 9, pp. 171–185, 1995.

[13] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002, vol. 1, pp. 557–600.

[14] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with rasta-plp," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, vol. 2, pp. 1259–1262.

[15] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Speech Audio Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[16] D. Gesbert and P. Duhamel, "Robust blind identification and equalization based on multi-step predictors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, vol. 26(5), pp. 3621–3624.

[17] B. W. Gillespie, H. S. Malvar, and D. A. F. Florncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2001, vol. 1, pp. 3701–3704.

[18] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using lp residual," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[19] A.Álvarez, V. Nieto, P. Gómez, and R. Martínez, "Speech enhancement based on linear prediction error signals and spectral subtraction," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2003, vol. 1, pp. 123–126.

[20] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2003, vol. 1, pp. 99–102.

[21] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ: Prentice-Hall, 2000.

[22] D. A. Harville, *Matrix Algebra from a Statistician's Perspective*. New York: Springer, 1997.

[23] M. Delcroix, T. Hikichi, and M. Miyoshi, "Blind dereverberation algorithm for speech signals based on multi-channel linear prediction," *Acoust. Sci. Technol.*, vol. 26, no. 5, pp. 432–439, 2005.

[24] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[25] J. B. Allen and D. A. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[26] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step linear prediction for single channel speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, vol. 1, pp. 817–820.

[27] J. J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 1995.

[28] D. Kressner and P. V. Dooren, "Factorizations and linear system solvers for matrices with toeplitz structure – SLICOT Working Note," Tech. Rep. TU Berlin, Berlin, Germany, 2000.

[29] A. Varga and P. Benner, "SLICOT – A subroutine library in systems and control theory," *Appl. Comput. Control, Signal Circuits*, vol. 1, pp. 499–539, 1999.

[30] P. Bondon, P. D. Ruiz, and A. Gallego, "Recursive methods for estimating multiple missing values of amultivariate stationary process," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1998, vol. 3, pp. 1361–1364.

[31] N. Kitawaki, M. Honda, and K. Itoh, "Speech-quality assessment methods for speech-coding systems," *IEEE Commun. Mag.*, vol. 22, no. 10, pp. 26–33, 1984.

[32] H. Kuttruff, *Room Acoustics*. New York: Spon Press, 2000.

[33] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, Sep. 1996.

[34] F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models," in *Proc. Eurospeech*, 1993, pp. 1031–1034.

[35] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Multi-step linear prediction based speech dereverberation in noisy reverberant environment," in *Proc. Interspeech*, 2007, pp. 854–857.

Keisuke Kinoshita (M'05) received the M.E. degree from Sophia University, Tokyo, Japan, in 2003.

He is currently a Member of Research Staff at NTT Communication Science Laboratories, NTT Corporation, and is engaged in research on speech and music signal processing.

Mr. Kinoshita was honored to receive the 2004 ASJ Poster Award, the 2004 ASJ Kansai Young Researcher Award, and the 2005 IEICE Best Paper Award. He is a member of the ASJ and IEICE.

Marc Delcroix (M'07) was born in Brussels, Belgium, in 1980. He received the M.Eng. degree from the Free University of Brussels and the Ecole Centrale Paris in 2003 and the Ph.D. degree from the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan, in 2007.

From 2004 to 2008, he was a Researcher at NTT Communication Science Laboratories, Kyoto, Japan, and worked on speech dereverberation and speech recognition. He is now with Pixela Corporation, Osaka, Japan, on software development for digital television.

Dr. Delcroix received the 2005 Young Researcher Awards from the Kansai Section of the Acoustic Society of Japan, the 2006 Student Paper Awards from the IEEE Kansai Section, and the 2006 Sato Paper Awards from the ASJ.

Tomohiro Nakatani (SM'06) received the B.E., M.E., and Ph.D. degrees from Kyoto University, Kyoto, Japan in 1989, 1991, and 2002, respectively.

He is a Senior Research Scientist with NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. Since he joined NTT Corporation as a Researcher in 1991, he has been investigating speech enhancement technologies for developing intelligent human–machine interfaces. From 1998 to 2001, he was engaged in developing multimedia services at business departments of NTT and NTT-East Corporations. In 2005, he visited the Georgia Institute of Technology, Atlanta, as a Visiting Scholar for a year.

Dr. Nakatani was honored to receive the 1997 JSAI Conference Best Paper Award, the 2002 ASJ Poster Award, and the 2005 IEICE Paper Awards. He is a member of the IEEE CAS Blind Signal Processing Technical Committee, an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and a Technical Program Chair of IEEE WASPAA-2007. He is a member of the IEICE and ASJ.

Masato Miyoshi (SM'04) received the M.E. and Ph.D. degrees from Doshisha Univerity, Kyoto, Japan, in 1983 and 1991, respectively.

Since joining NTT Corproation , Kyoto, Japan, as a Researcher in 1983, he has been studying signal processing theory and its application to acoustic technologies. Currently, he is the leader of the Signal Processing Group, the Media Information Lab, NTT Communication Science Laboratories. He is also a Guest Professor of the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan.

Dr. Miyoshi was honored to receive the 1988 IEEE ASSP Society's Senior Award, the 1989 ASJ Kiyoshi-Awaya Incentive Award, the 1990 and 2006 ASJ Sato Paper Awards, and the 2005 IEICE Paper Award. He is a member of the IEICE, ASJ, and AES.