Integrated Speech Enhancement Method Using Noise Suppression and Dereverberation

Takuya Yoshioka, *Member, IEEE*, Tomohiro Nakatani, *Senior Member, IEEE*, and Masato Miyoshi, *Senior Member, IEEE*

***Abstract*—This paper proposes a method for enhancing speech signals contaminated by room reverberation and additive stationary noise. The following conditions are assumed. 1) Short-time spectral components of speech and noise are statistically independent Gaussian random variables. 2) A room's convolutive system is modeled as an autoregressive system in each frequency band. 3) A short-time power spectral density of speech is modeled as an all-pole spectrum, while that of noise is assumed to be time-invariant and known in advance. Under these conditions, the proposed method estimates the parameters of the convolutive system and those of the all-pole speech model based on the maximum likelihood estimation method. The estimated parameters are then used to calculate the minimum mean square error estimates of the speech spectral components. The proposed method has two significant features. 1) The parameter estimation part performs noise suppression and dereverberation alternately. (2) Noise-free reverberant speech spectrum estimates, which are transferred by the noise suppression process to the dereverberation process, are represented in the form of a probability distribution. This paper reports the experimental results of 1500 trials conducted using 500 different utterances. The reverberation time RT_{60} was 0.6 s, and the reverberant signal to noise ratio was 20, 15, or 10 dB. The experimental results show the superiority of the proposed method over the sequential performance of the noise suppression and dereverberation processes.**

***Index Terms*—Dereverberation, maximum-likelihood (ML) estimation, minimum mean square error (MMSE) estimation, noise suppression, speech enhancement.**

I. Introduction

SPEECH signals captured by microphones in rooms are often corrupted by both reverberation and background noise. The recovery of the clean speech signals from the observed noisy reverberant signals will be indispensable for many audio applications.

Fig. 1 shows the acoustic system of interest. A clean speech signal is first reverberated by a room's linear convolutive system; then, the reverberant speech signal is corrupted by an additive noise signal at a microphone. In this paper, we assume that only one microphone is available.

Generally speaking, the following three problems are involved in the enhancement of noisy reverberant speech signals:

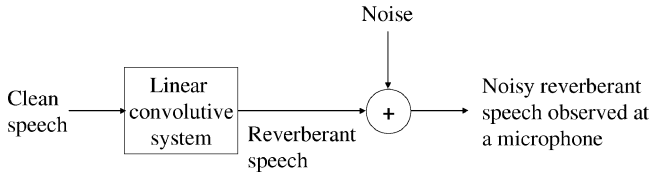

Fig. 1. Acoustic system of interest.

1) noise suppression;
2) dereverberation;
3) integration of noise suppression and dereverberation processes.

The noise suppression task has been investigated intensively for decades. We refer the reader to [1, Ch. 13] for individual methods. One useful insight is that many conventional noise suppression methods are based on statistical models of speech. For example, [2]–[4] use an all-pole model of speech.

On the other hand, in recent years, the speech dereverberation task has attracted a lot of attention. The subspace method is a commonly used dereverberation approach [5], [6]. However, it is extremely sensitive to additive noise [7], [8]. Prewhitening-based methods are relatively robust as regards additive noise [9]–[11]. Other methods include harmonicity-based [12], sparsity-based [13], nonstationarity-based [14], and common-pole-detection-based [15] techniques. In a series of recent publications, we have proposed several speech dereverberation methods based on statistical speech models [16]–[20]. The all-pole speech model is used in [16] and [17], whereas a speech autocorrelation codebook is used in [18]. These methods exhibit robustness with respect to relatively low magnitude noise.

The third problem, namely the integration of noise suppression and dereverberation processes has yet to be investigated. As far as we know, only two papers deal with related issues [21], [22].[1] The method in [21] seems incapable of cancelling long reverberation. [22] proposes performing the noise suppression and dereverberation processes sequentially. To be more precise, the method in [22] suppresses noise components by spectral subtraction [24], which is followed by applying the dereverberation method in [13] to the noise-suppressed signal. However, we believe that the noise suppression and dereverberation processes should cooperate with each other to achieve higher speech enhancement performance.

This paper proposes a speech enhancement method that effectively integrates the noise suppression and dereverberation

Manuscript received March 26, 2008; revised August 27, 2008. Current version published February 11, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tim Fingscheidt.

The authors are with the NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kyoto 619-0237, Japan (e-mail: takuya@cslab.kecl.ntt.co.jp; nak@cslab.kecl.ntt.co.jp; miyo@cslab.kecl.ntt.co.jp).

Digital Object Identifier 10.1109/TASL.2008.2008042

[1]Although [17] and [23] also address the noise suppression and dereverberation, these papers assume that there are multiple microphones and that the noise is generated by a point sound source.

1558-7916/$25.00 © 2009 IEEE

processes. For the sake of simplicity, the power spectral density (PSD) of the noise is assumed to be known in advance. In practice, the noise PSD can be obtained from the observed signal during speech pauses or by using other noise estimation methods such as [4] and [25].

The proposed method is based on a statistical approach because such approaches provide satisfactory results in both the noise suppression task and the dereverberation task. The proposed method consists of a parameter estimation part and a spectrum estimation part. The parameter estimation part estimates the parameters of an assumed model from the short-time spectra of an observed noisy reverberant speech signal by using the maximum-likelihood (ML) estimation method. The spectrum estimation part calculates minimum mean square error (MMSE) estimates of clean speech spectra by using the estimated model parameters.

The model parameters to be estimated are those of a short-time speech spectrum and those of a room's convolutive system. For the speech model, the all-pole model is used, which has been successfully exploited for both noise suppression and dereverberation as described above. On the other hand, the room's convolutive system is modeled as an autoregressive (AR) system in each frequency band. This convolutive system model was proposed in [20]. We will discuss the merits and demerits of this model in Section III-A. One advantage of this model facilitates the integration of noise suppression and dereverberation because most conventional noise suppression methods are based on the frequency-domain signal representation.

Significant points for emphasis are as follows.

- The parameter estimation part performs noise suppression and dereverberation processes alternately. This contrasts with the method in [22], which performs these two processes sequentially.
- Noise-free reverberant speech spectrum estimates, which are transferred by the noise suppression process to the dereverberation process, are represented in the form of a probability distribution. This allows the dereverberation process to take account of the uncertainty of the noise-free reverberant speech spectrum estimates.

This paper is organized as follows: Section II defines the noisy reverberant speech enhancement task. Section III describes the statistical model of observed noisy reverberant speech spectra. Sections IV and V explain the spectrum estimation part and the parameter estimation part, respectively. Section VI introduces some modifications. Section VII reports experimental results, and Section VIII concludes this paper.

II. Task Description of Noisy Reverberant Speech Enhancement

In this section, we define the task of noisy reverberant speech enhancement. Let $s(n)$ and $d(n)$ denote a clean speech signal and a noise signal, respectively, where n is the time index. In addition, let $x(n)$ and $y(n)$ be the corresponding noise-free reverberant speech signal and noisy reverberant speech signal, respectively. $x(n)$ and $y(n)$ are generated as

$$x(n) = \sum_{k=0}^{\infty} h(k)s(n-k) \tag{1}$$

$$y(n) = x(n) + d(n) \tag{2}$$

respectively, where $\{h(k)\}_{0\le k\le\infty}$ is the impulse response of the linear convolutive system from the speaker to the microphone, which is often called a room impulse response.

Now, suppose that we observe N samples of the noisy reverberant signal, which is denoted by

$$\mathcal{Y}_{\text{time}} = \{y(n)\}_{0\le n\le N-1}. \tag{3}$$

Then, the task of noisy reverberant speech enhancement is defined as estimating the corresponding samples of the clean speech signal

$$\mathcal{S}_{\text{time}} = \{s(n)\}_{0\le n\le N-1}. \tag{4}$$

First, we have to choose a domain where an observed signal, which is provided with a speech enhancement system as its input, is represented. The proposed method postulates that the observed signal is expressed in the time–frequency domain. The advantage of using the time–frequency representation will be discussed in Section III-A.

The time–frequency analysis is performed using a short-time Fourier transform (STFT) with an L-point frame length and a W-point frame shift. Let $S_{t,l}$ and $Y_{t,l}$ denote the spectral components of $s(n)$ and $y(n)$, respectively, at the tth frame and the lth frequency band. Note that the number of frequency bands is also L. Also let $\mathcal{S}$ and $\mathcal{Y}$ be the set of all clean speech spectral components corresponding to $\mathcal{S}_{\text{time}}$ and the set of all noisy reverberant speech spectral components corresponding to $\mathcal{Y}_{\text{time}}$, respectively. If we use T to denote the number of frames corresponding to the observed sample number N, $\mathcal{S}$, and $\mathcal{Y}$ are given, respectively, by

$$\mathcal{S} = \{S_{t,l}\}_{0\le t\le T-1, 0\le l\le L-1} \tag{5}$$

$$\mathcal{Y} = \{Y_{t,l}\}_{0\le t\le T-1, 0\le l\le L-1}. \tag{6}$$

The speech enhancement task is now redefined as follows.

> Estimate the spectral components of the clean speech signal $\mathcal{S}$ from those of the noisy reverberant speech signal $\mathcal{Y}$.

The enhanced time-domain clean speech signal $\hat{s}(n)$ is synthesized from the estimated clean speech spectral components by using the overlap-add synthesis technique.

III. Modeling of Noisy Reverberant Speech

Solving the above estimation task requires a criterion whereby we measure the degree of appropriateness of clean speech spectral component estimates. Establishing a statistical

model of the spectral components of noisy reverberant speech is a commonly used way of introducing such a criterion. It is natural for the noisy reverberant speech model to be composed of models of reverberation (or a room's convolutive system), speech, and noise. Therefore, we describe the reverberation model in Section III-A, and speech and noise models in Section III-B. Subsequently, we incorporate these models in the noisy reverberant speech model in Section III-C.

A. Reverberation Model

A noise-free reverberant speech signal $x(n)$ is related to a clean speech signal $s(n)$ by (1) in the time domain. By contrast, our goal is to estimate the spectral components of the clean speech signal, which are defined in the time–frequency domain. Hence, it is desirable to associate reverberant speech spectral components with clean speech spectral components directly in the time–frequency domain. In this paper, this association is accomplished as follows.

Let $X_{t,l}$ denote the spectral component of $x(n)$ at the tth frame and the lth frequency band. In this paper, we assume that the sequence of the reverberant speech spectral components in the lth frequency band $\{X_{t,l}\}_{0\leq t\leq T-1}$ is the output of an AR system driven by the sequence of the lth frequency-band clean speech spectral components, $\{S_{t,l}\}_{0\leq t\leq T-1}$. Thus, we have

$$X_{t,l} = \sum_{k=1}^{K_l} g_{k,l}^* X_{t-k,l} + S_{t,l} \tag{7}$$

where $g_{k,l}$ is the kth regression coefficient for the lth frequency band, K_l is the regression order, and superscript $*$ stands for the complex conjugate. We hereafter call $g_{k,l}$ a room regression coefficient.

As long as reverberation model (7) closely approximates time-domain reverberation model (1), each clean speech spectral component can be recovered by

$$S_{t,l} = X_{t,l} - \sum_{k=1}^{K_l} g_{k,l}^* X_{t-k,l}. \tag{8}$$

Equation (8) indicates that we can perform dereverberation by applying causal FIR filters to individual frequency band-wise sequences.

Reverberation model (7), or equivalently dereverberation model (8), enjoys the following advantages.

- The dereverberation system (8) is easily combined with a noise suppression system because many conventional noise suppression methods are based on the time–frequency domain signal representation.
- The filter length for each frequency band is reduced compared with the length of a time-domain dereverberation filter. The filter-length reduction contributes directly to the reduction of computational cost [20].
- In theory, the room's convolutive system is invertible with a causal FIR filter in the time-domain only if the system is minimum phase [26]. However, it was reported that clean speech spectral components may be well recovered with causal FIR filters in the time–frequency domain even when the room's convolutive system is non-minimum phase in the time-domain [27].
- Parameters associated with each frequency band are optimized independently of those associated with any other frequency bands. This enables us to implement dereverberation in a parallel processing manner.

Note however that reverberation model (7) has the following potential disadvantages.

- Model (7) indicates that there are no cross-band components. In other words, each reverberant speech spectral component in an arbitrary frequency band is assumed to be independent of spectral components in any other frequency bands. However, there should be some cross-band components if we are to make the time–frequency domain reverberation model equivalent to the time-domain reverberation model [28].
- Model (7) may be unable to represent the early reflection part, which is shorter than the frame shift W of the room's convolutive system because the zeroth tap weight of dereverberation filter (8) is fixed at 1 for all l values. This means that (8) is incapable of cancelling out the distortion caused by the early reflections. This is easily confirmed by considering a case where the room impulse response, $\{h(k)\}_{0\leq k\leq\infty}$ in (1), is shorter than W. In such a case, the room regression coefficient $g_{k,l}$ must be nearly equal to zero for any k and l, which leads to $S_{t,l} \simeq X_{t,l}$. Nevertheless, $S_{t,l}$ value may differ greatly from $X_{t,l}$ value due to the room impulse response. Note however that the early reflections have no destructive effects on the speech quality or automatic speech recognition performance [13], [29].

The time–frequency domain AR modeling of the room's convolutive system was experimentally confirmed to be effective for the speech dereverberation task in [20] and [30]. The model has also been discussed theoretically in [20] especially with respect to multiple microphone systems.

B. Speech and Noise Models

Next, let us define the speech model. We assume that a short-time spectrum of clean speech is represented by an all-pole model. Specifically, the clean speech spectrum is assumed to satisfy the following conditions.

1) The power spectral density (PSD) of clean speech is of the all-pole form [31]. Thus, by letting ${}_s\lambda_t(\omega)$ denote the PSD at the tth frame and normalized angular frequency ω, we have

$$\begin{aligned} {}_s\lambda_t\left(\omega; \boldsymbol{a}_t, {}_s\sigma_t^2\right) &= \frac{{}_s\sigma_t^2}{\left|A_t(e^{j\omega}; \boldsymbol{a}_t)\right|^2} \\ A_t(z; \boldsymbol{a}_t) &= 1 - a_{t,1}z^{-1} - \cdots - a_{t,P}z^{-P} \\ \boldsymbol{a}_t &= [a_{t,1}, \cdots, a_{t,P}]^\top \end{aligned} \tag{9}$$

where P is the number of poles, $\{\boldsymbol{a}_t, {}_s\sigma_t^2\}$ is the set of all-pole model parameters for the tth frame, and superscript $\top$ denotes non-conjugate transposition. $\boldsymbol{a}_t$ and ${}_s\sigma_t^2$ are called linear prediction coefficients (LPCs) and a prediction residual, respectively. Hereafter, the description of the parameters will be omitted for brevity unless otherwise

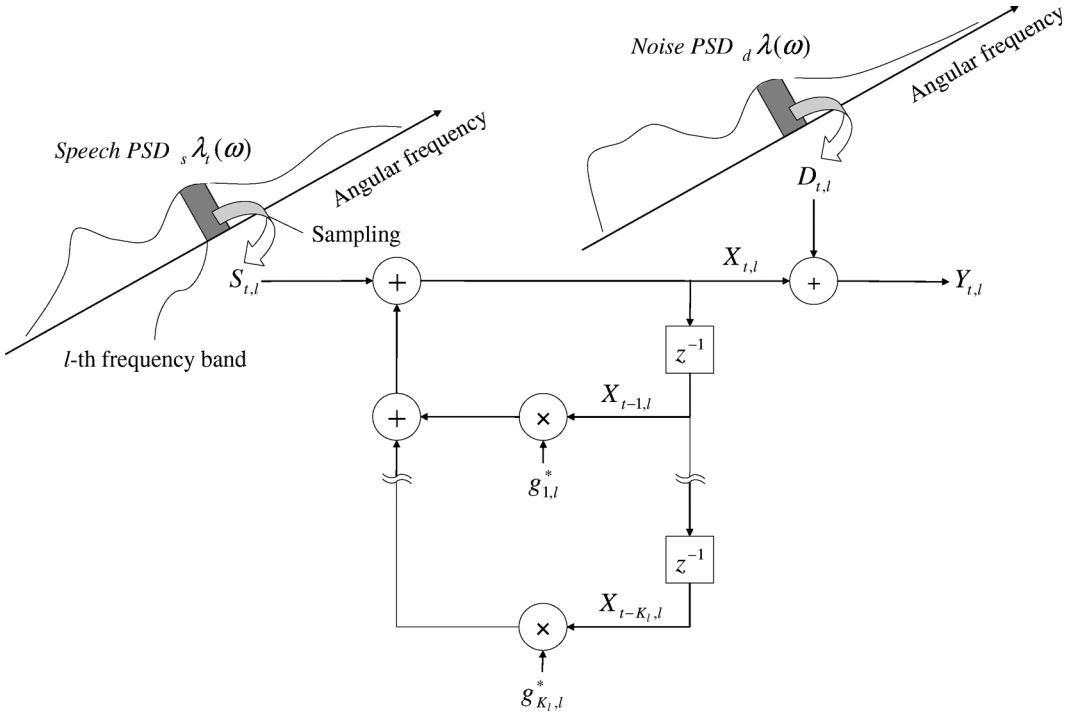

Fig. 2. Schematic diagram of generative model for noisy reverberant speech.

noted; for example, ${}_s\lambda_t(\omega; \boldsymbol{a}_t, {}_s\sigma_t^2)$ is written as ${}_s\lambda_t(\omega)$ for short.

2) Each spectral component $S_{t,l}$ of clean speech follows a complex Gaussian process [32] with mean 0 and variance ${}_s\lambda_t(2\pi l/L)$ [33]. Note that $\omega = 2\pi l/L$ is the normalized angular frequency corresponding to the lth frequency band. Hence, the probability density function (pdf) of $S_{t,l}$ is

$$p\left(S_{t,l}; \boldsymbol{a}_t, {}_s\sigma_t^2\right) = \mathcal{N}_{\mathbb{C}}\left\{S_{t,l}; 0, {}_s\lambda_t\left(\frac{2\pi l}{L}\right)\right\} = \frac{1}{\pi\, {}_s\lambda_t(2\pi l/L)} \exp\left\{-\frac{|S_{t,l}|^2}{{}_s\lambda_t(2\pi l/L)}\right\}. \quad (10)$$

$\mathcal{N}_{\mathbb{C}}\{\boldsymbol{x}; \boldsymbol{\mu}, \Sigma\}$ stands for the pdf of arbitrary complex Gaussian random variable $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ, which is defined as

$$\mathcal{N}_{\mathbb{C}}\{\boldsymbol{x}; \boldsymbol{\mu}, \Sigma\} = \frac{1}{\pi^C \det \Sigma} \exp\left\{(\boldsymbol{x} - \boldsymbol{\mu})^H \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right\} \quad (11)$$

where C is the dimension of $\boldsymbol{x}$. Especially when $\boldsymbol{x}$ is of one dimension and rewritten as x to highlight this, the pdf is represented as

$$\mathcal{N}_{\mathbb{C}}\{x; \mu, \sigma^2\} = \frac{1}{\pi\sigma^2} \exp\left\{-\frac{|x - \mu|^2}{\sigma^2}\right\} \quad (12)$$

where μ and σ^2 are the mean and variance of x, respectively.

3) If $(t, l) \neq (t', l')$, $S_{t,l}$ and $S_{t',l'}$ are statistically independent [33].

These assumptions have been widely accepted in the literature [2], [3].

Finally, noise signals are assumed to meet the following conditions. Below, we represent the spectral component of $d(n)$ at the tth frame and the lth frequency band by $D_{t,l}$.

1) A noise signal is stationary; in other words, the PSD of the noise signal is time-invariant. Hereafter, we denote the noise PSD by ${}_d\lambda(\omega)$. Note that ${}_d\lambda(\omega)$ is independent of frame index t. Moreover, the noise PSD is assumed to be known in advance, and may be estimated from observation intervals where speech is absent.
2) Each noise spectral component $D_{t,l}$ follows a complex Gaussian process with mean 0 and variance ${}_d\lambda(2\pi l/L)$

$$p\left(D_{t,l}; {}_d\lambda\left(\frac{2\pi l}{L}\right)\right) = \mathcal{N}_{\mathbb{C}}\left\{D_{t,l}; 0, {}_d\lambda\left(\frac{2\pi l}{L}\right)\right\} = \frac{1}{\pi_d \lambda(2\pi l/L)} \times \exp\left\{-\frac{|D_{t,l}|^2}{{}_d\lambda(2\pi l/L)}\right\}. \quad (13)$$

3) If $(t, l) \neq (t', l')$, $D_{t,l}$ and $D_{t',l'}$ are statistically independent. In addition, $S_{t,l}$ and $D_{t',l'}$ are statistically independent for any (t, l, t', l').

C. Model of Noisy Reverberant Speech

Now that the models of reverberation, speech, and noise have been defined, we can obtain a statistical model of noisy reverberant speech spectral components. Based on the reverberation model, the model for generating the noisy reverberant speech can be depicted as Fig. 2. The joint pdf of noisy reverberant speech spectral components $\mathcal{Y}$ and noise-free reverberant

speech spectral components $\mathcal{X} = \{X_{t,l}\}_{0\leq t\leq T-1, 0\leq l\leq L-1}$ can be described as

$$\begin{aligned} &p(\mathcal{Y},\mathcal{X};\Theta) \\ &\propto \left(\prod_{l=0}^{L-1} {}_d\lambda\left(\frac{2\pi l}{L}\right)^{-T}\right)\left(\prod_{t=0}^{T-1}\prod_{l=0}^{L-1}\frac{\left|A_t\left(e^{j\frac{2\pi l}{L}}\right)\right|^2}{{}_s\sigma_t^2}\right) \\ &\quad\times\exp\left\{-\sum_{t=0}^{T-1}\sum_{l=0}^{L-1}\left(\frac{|Y_{t,l}-X_{t,l}|^2}{{}_d\lambda(2\pi l/L)}+\frac{\left|A_t\left(e^{j\frac{2\pi l}{L}}\right)\right|^2}{{}_s\sigma_t^2}\right.\right. \\ &\qquad\left.\left.\times\left|X_{t,l}-\sum_{k=1}^{K_l}g_{k,l}^* X_{t-k,l}\right|^2\right)\right\} \end{aligned} \tag{14}$$

where Θ is the set of all parameters [see Appendix A for the derivation of (14)]. Specifically, Θ is defined as

$$\Theta = \{{}_s\Theta, {}_g\Theta, {}_d\Theta\} \tag{15}$$

$${}_s\Theta = \left\{\boldsymbol{a}_t, {}_s\sigma_t^2\right\}_{0\leq t\leq T-1} \tag{16}$$

$${}_g\Theta = \{g_{1,l},\cdots,g_{K_l,l}\}_{0\leq l\leq L-1} \tag{17}$$

$${}_d\Theta = \left\{{}_d\lambda\left(\frac{2\pi l}{L}\right)\right\}_{0\leq l\leq L-1}. \tag{18}$$

Here, ${}_s\Theta$, ${}_g\Theta$, and ${}_d\Theta$ are sets of speech parameters, room regression coefficients, and noise parameters, respectively.

The joint pdf (14) can be simplified by using the Szegö's theorem [34]. The theorem leads to

$$\lim_{L\to\infty}\sum_{l=0}^{L-1}\log A_t\left(e^{j\frac{2\pi l}{L}}\right)=0. \tag{19}$$

Provided that the number of frames L is sufficiently large, substituting (19) into (14) yields an approximated form of the joint pdf as

$$\begin{aligned} &p(\mathcal{Y},\mathcal{X};\Theta) \\ &\propto \left(\prod_{l=0}^{L-1} {}_d\lambda\left(\frac{2\pi l}{L}\right)^{-T}\right)\left(\prod_{t=0}^{T-1}\left({}_s\sigma_t^2\right)^{-L}\right) \\ &\quad\times\exp\left\{-\sum_{t=0}^{T-1}\sum_{l=0}^{L-1}\left(\frac{|Y_{t,l}-X_{t,l}|^2}{{}_d\lambda(2\pi l/L)}+\frac{\left|A_t\left(e^{j\frac{2\pi l}{L}}\right)\right|^2}{{}_s\sigma_t^2}\right.\right. \\ &\qquad\left.\left.\times\left|X_{t,l}-\sum_{k=1}^{K_l}g_{k,l}^* X_{t-k,l}\right|^2\right)\right\}. \end{aligned} \tag{20}$$

With (20), the statistical model of noisy reverberant speech spectral components has been defined.

IV. Minimum Mean Square Error Estimation of Clean Speech

For the time being, let us suppose that parameter set Θ is known in advance. In this section, we derive the minimum mean square error (MMSE) estimate of clean speech spectral component $S_{t,l}$, which is given by

$$\hat{S}_{t,l} = \langle S_{t,l}\rangle_{p(S_{t,l}|\mathcal{Y};\Theta)}. \tag{21}$$

Notation $\langle f(x)\rangle_{p(x)}$ means the expected value of $f(x)$ over random variable x with pdf $p(x)$.

First, we derive the conditional posterior, $p(\mathcal{X}|\mathcal{Y};\Theta)$, of reverberant speech spectral components $\mathcal{X}$ given noisy reverberant speech spectral components $\mathcal{Y}$ as follows. Let us put all the lth frequency-band spectral components of reverberant speech and those of noisy reverberant speech, respectively, in the following vectors:

$$\boldsymbol{X}_l = [X_{T-1,l}, X_{T-2,l},\cdots,X_{0,l}]^\top \tag{22}$$

$$\boldsymbol{Y}_l = [Y_{T-1,l}, Y_{T-2,l},\cdots,Y_{0,l}]^\top. \tag{23}$$

Note that $\mathcal{X}$ and $\mathcal{Y}$ correspond, respectively, to the sets of $\boldsymbol{X}_l$ and $\boldsymbol{Y}_l$ over all the frequency bands. Then, the conditional posterior $p(\mathcal{X}|\mathcal{Y};\Theta)$ is represented as

$$p(\mathcal{X}|\mathcal{Y};\Theta) = \prod_{l=0}^{L-1}\mathcal{N}_{\mathbb{C}}\left\{\boldsymbol{X}_l;\boldsymbol{\mu}_l(\Theta,\boldsymbol{Y}_l),\Sigma_l(\Theta)\right\}. \tag{24}$$

Mean $\boldsymbol{\mu}_l(\Theta,\boldsymbol{Y}_l)$ and covariance matrix $\Sigma_l(\Theta)$ are given, respectively, by

$$\begin{aligned} \boldsymbol{\mu}_l(\Theta,\boldsymbol{Y}_l) = &\left\{B_l({}_d\Theta)B_l({}_d\Theta)^H\right. \\ &\left.+G_l({}_g\Theta)A_l({}_s\Theta)A_l({}_s\Theta)^H G_l({}_g\Theta)^H\right\}^{-1} \\ &\times\left\{B_l({}_d\Theta)B_l({}_d\Theta)^H\right\}\boldsymbol{Y}_l \end{aligned} \tag{25}$$

$$\begin{aligned} \Sigma_l(\Theta) = &\left\{B_l({}_d\Theta)B_l({}_d\Theta)^H + G_l({}_g\Theta)A_l({}_s\Theta)\right. \\ &\left.\times A_l({}_s\Theta)^H G_l({}_g\Theta)^H\right\}^{-1} \end{aligned} \tag{26}$$

where superscript H denotes conjugate transposition. $G_l({}_g\Theta)$, $A_l({}_s\Theta)$, and $B_l({}_d\Theta)$ are T-dimensional square matrices defined, respectively, as

$$G_l({}_g\Theta) = \begin{bmatrix} 1 & & & & & \\ -g_{1,l} & \ddots & & & & \\ \vdots & & 1 & & & \\ -g_{K_l,l} & & -g_{1,l} & \ddots & & \\ & \ddots & \vdots & & \ddots & \\ & & -g_{K_l,l} & \cdots & \cdots & 1 \end{bmatrix} \tag{27}$$

$$A_l({}_s\Theta) = \mathrm{diag}\left\{{}_s\lambda_{T-1}\left(\frac{2\pi l}{L}\right)^{-\frac{1}{2}},\cdots,{}_s\lambda_0\left(\frac{2\pi l}{L}\right)^{-\frac{1}{2}}\right\} \tag{28}$$

$$B_l({}_d\Theta) = \mathrm{diag}\left\{{}_d\lambda\left(\frac{2\pi l}{L}\right)^{-\frac{1}{2}},\cdots,{}_d\lambda\left(\frac{2\pi l}{L}\right)^{-\frac{1}{2}}\right\} \tag{29}$$

where $\mathrm{diag}\{c_1,\cdots,c_C\}$ represents the diagonal matrix with c_m on its mth diagonal element. The derivation of (24) is provided in Appendix B. For brevity, we will hereafter omit the arguments of the above variables unless otherwise noted; for example, $\boldsymbol{\mu}_l(\Theta,\boldsymbol{Y}_l)$ is written as $\boldsymbol{\mu}_l$ for short.

It is noteworthy that (25) is an extension of a well-known Wiener filter-based noise suppression technique [1, Sec. 13.3].

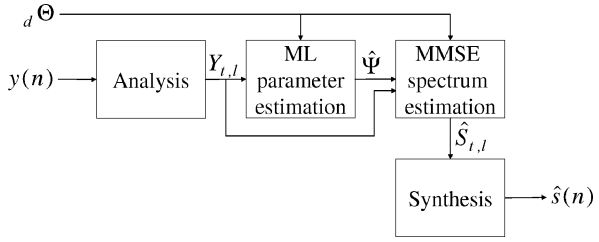

Fig. 3. Processing flow of the proposed speech enhancement system.

This can be easily confirmed by letting $g_{k,l} = 0$ for any k and l in (25), namely, by ignoring the presence of reverberation.

By using $p(\mathcal{X}|\mathcal{Y};\Theta)$, the MMSE estimate of $S_{t,l}$, namely $\hat{S}_{t,l}$, is derived as follows. Let us define $\boldsymbol{S}_l$ as

$$\boldsymbol{S}_l = [S_{T-1,l}, \cdots, S_{0,l}]^\top. \tag{30}$$

$\boldsymbol{S}_l$ is associated with $\boldsymbol{X}_l$ by

$$\boldsymbol{S}_l = G_l^H \boldsymbol{X}_l. \tag{31}$$

Therefore, from (24), we obtain the conditional posterior of clean speech spectral components

$$p(\mathcal{S}|\mathcal{Y};\Theta) = \prod_{l=0}^{L-1} \mathcal{N}_{\mathbb{C}}\left\{\boldsymbol{S}_l; G_l^H \boldsymbol{\mu}_l, G_l^H \Sigma_l G_l\right\}. \tag{32}$$

Hence, the MMSE estimate of $S_{t,l}$, or $\hat{S}_{t,l}$, is given as the $(T-t)$th element of $G_l^H \boldsymbol{\mu}_l$. If we write the $(T-t)$th element of $\boldsymbol{\mu}_l$ by $\mu_{t,l}$, i.e.,

$$\boldsymbol{\mu}_l = [\mu_{T-1,l}, \cdots, \mu_{0,l}]^\top \tag{33}$$

we obtain

$$\hat{S}_{t,l} = \mu_{t,l} - \sum_{k=1}^{K_l} g_{k,l}^* \mu_{t-k,l}. \tag{34}$$

In reality, since parameter set Θ is unknown in advance, we have to estimate Θ from observed data $\mathcal{Y}$. As mentioned previously, noise parameters $_d\Theta$ included in Θ are assumed to be given beforehand. Therefore, we only have to estimate the remaining parameters, which we denote by

$$\Psi = \{_s\Theta, _g\Theta\}. \tag{35}$$

Hence, a speech enhancement system based on the proposed method is structured as shown in Fig. 3. We will explain the proposed algorithm for estimating Ψ in the next section.

V. Maximum-Likelihood Parameter Estimation

To estimate unknown parameters Ψ, we employ the ML estimation method. With the ML estimation method, once noisy reverberant speech spectral components $\mathcal{Y}$ are observed, $\hat{\Psi}$ that (locally) maximizes likelihood function $p(\mathcal{Y};\Psi,{}_d\Theta)$ are calculated as the estimates of Ψ. In this section, we describe an algorithm for calculating $\hat{\Psi}$.

A. Expectation-Conditional Maximization Algorithm

Noisy reverberant speech model (20) involves reverberant speech spectral components $\mathcal{X}$ as latent variables. Therefore, we shall resort to the expectation maximization (EM) algorithm or its variants for calculating ML estimates $\hat{\Psi}$.

The EM algorithm is based on iterative processing. Observable variables $\mathcal{Y}$ and latent variables $\mathcal{X}$ are collectively called complete variables. Let $\hat{\Psi}^{(i)}$ denote the tentative estimates of Ψ after the ith iteration is completed. Once $\hat{\Psi}^{(i)}$ is given, the expected log likelihood of Ψ can be defined as

$$Q\left(\Psi;\hat{\Psi}^{(i)}\right) = \left\langle \log p(\mathcal{Y},\mathcal{X};\Psi,{}_d\Theta)\right\rangle_{p\left(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta\right)}. \tag{36}$$

$Q(\Psi;\Psi^{(i)})$ is usually called an auxiliary function. In each iteration, the EM algorithm maximizes the auxiliary function with respect to Ψ according to the following two steps.

EM algorithm

1) *E-step*
   Calculate the conditional posterior, $p(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta)$, of reverberant speech spectral components given tentative parameter estimates $\hat{\Psi}^{(i)}$.
2) *M-step*
   Update the estimates of Ψ by
$$\hat{\Psi}^{(i+1)} = \arg\max_{\Psi} Q\left(\Psi;\Psi^{(i)}\right). \tag{37}$$
   Then, increment iteration index i by 1.

Unfortunately, however, there is no closed-form solution for maximization step (37). Because Ψ is the union of $_s\Theta$ and $_g\Theta$ as (35), we split M-step into two steps: the first step maximizes the auxiliary function with respect to $_s\Theta$ for fixed $_g\Theta$. On the other hand, the second step maximizes the auxiliary function with respect to $_g\Theta$ while keeping $_s\Theta$ invariant. As shown later, both maximization steps are analytically solvable. This modified EM algorithm is called the expectation-conditional maximization (ECM) algorithm and is proven to converge to a local optimum [35]. Each iteration of the ECM algorithm consists of the following three steps.

ECM Algorithm

1) *E-step*
   Calculate the posterior, $p(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta)$, of reverberant speech spectral components given tentative parameter estimates $\hat{\Psi}^{(i)}$.
2) *CM-step1*
   Update the estimates of $_s\Theta$ by
$$_s\hat{\Theta}^{(i+1)} = \arg\max_{{}_s\Theta} Q\left({}_s\Theta, {}_g\hat{\Theta}^{(i)};\hat{\Psi}^{(i)}\right). \tag{38}$$
3) *CM-step2*
   Update the estimates of $_g\Theta$ by
$$_g\hat{\Theta}^{(i+1)} = \arg\max_{{}_g\Theta} Q\left({}_s\hat{\Theta}^{(i+1)}, {}_g\Theta;\hat{\Psi}^{(i)}\right). \tag{39}$$
   Then, increment iteration index i by 1.

Below, we will derive the formulas for E-step, CM-step1, and CM-step2 in Sections V-B–V-D, respectively, and then provide a conceptual explanation of the derived algorithm.

B. E-step

E-step calculates the conditional posterior, $p(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta)$, of reverberant speech spectral components given tentative parameter estimates $\hat{\Psi}^{(i)}$. More specifically, E-step calculates the parameters specifying $p(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta)$. We have already derived $p(\mathcal{X}|\mathcal{Y};\Theta)$ as (24). By substituting $\Psi=\hat{\Psi}^{(i)}$ into (24), we obtain $p(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta)$ as

$$p\left(\mathcal{X}|\mathcal{Y},\hat{\Psi}^{(i)},{}_d\Theta\right)=\prod_{l=0}^{L-1}\mathcal{N}_{\mathbb{C}}\left\{\boldsymbol{X}_l;\boldsymbol{\mu}_l\left(\hat{\Psi}^{(i)},{}_d\Theta,\boldsymbol{Y}_l\right),\Sigma_l\left(\hat{\Psi}^{(i)},{}_d\Theta\right)\right\}. \quad (40)$$

Mean $\boldsymbol{\mu}_l(\hat{\Psi}^{(i)},{}_d\Theta,\boldsymbol{Y}_l)$ and covariance matrix $\Sigma_l(\hat{\Psi}^{(i)},{}_d\Theta)$ are obtained by letting $\Psi=\hat{\Psi}^{(i)}$ in (25) and (26), respectively.

Recall that the mean-calculation formula (25) $\boldsymbol{\mu}_l(\hat{\Psi}^{(i)},{}_d\Theta,\mathcal{Y})$, is an extension of the Wiener filter based noise suppression technique. Therefore, we may see that the role of E-step is to estimate noise-free reverberant speech spectral components from observed noisy reverberant speech spectral components.

For later use, we denote the mth element of $\boldsymbol{\mu}_l(\hat{\Psi}^{(i)},{}_d\Theta;\boldsymbol{Y}_l)$ by $\mu^{(i)}_{m,l}$, and the (m,n)th element of $\Sigma_l(\hat{\Psi}^{(i)},{}_d\Theta)$ by $\xi^{(i)}_{(m,n),l}$. In addition, we define subvector $\boldsymbol{\mu}^{(i)}_{m:n,l}$ of $\boldsymbol{\mu}_l(\hat{\Psi}^{(i)},{}_d\Theta;\boldsymbol{Y}_l)$ and submatrix $\Sigma^{(i)}_{(m:n,m':n'),l}$ of $\Sigma_l(\hat{\Psi}^{(i)},{}_d\Theta)$, respectively, as

$$\boldsymbol{\mu}^{(i)}_{m:n,l}=\left[\mu^{(i)}_{m,l},\cdots,\mu^{(i)}_{n,l}\right]^{\top}$$
$$\Sigma^{(i)}_{(m:n,m':n'),l}=\begin{bmatrix}\xi^{(i)}_{(m,n),l} & \cdots & \xi^{(i)}_{(m,n'),l}\\ \vdots & \ddots & \vdots\\ \xi^{(i)}_{(m',n),l} & \cdots & \xi^{(i)}_{(m',n'),l}\end{bmatrix}. \quad (41)$$

C. CM-step1

Next, we explain CM-step1, namely the update rules for estimates of speech parameters ${}_s\Theta$. Note that ${}_s\Theta$ coincides with the set of $\boldsymbol{a}_t$ and ${}_s\sigma^2_t$ over all the time frames.

Maximization step (38) is accomplished by updating the estimates of $\boldsymbol{a}_t$ and ${}_s\sigma^2_t$, for all frame indices t, according to the following rules:

$$\hat{\boldsymbol{a}}^{(i+1)}_t={}_sR^{(i)}_t{}^{-1}\,{}_s\boldsymbol{r}^{(i)}_t \quad (42)$$
$${}_s\hat{\sigma}^{2\,(i+1)}_t={}_sr^{(i)}_t(0)-{}_s\boldsymbol{r}^{(i)}_t{}^{\top}\,{}_sR^{(i)}_t{}^{-1}\,{}_s\boldsymbol{r}^{(i)}_t. \quad (43)$$

${}_sR^{(i)}_t$ and ${}_s\boldsymbol{r}^{(i)}_t$ in (42) and (43) are given, respectively, by

$${}_sR^{(i)}_t=\begin{bmatrix}{}_sr^{(i)}_t(0) & \cdots & {}_sr^{(i)}_t(P-1)\\ \vdots & \ddots & \vdots\\ {}_sr^{(i)}_t(P-1) & \cdots & {}_sr^{(i)}_t(0)\end{bmatrix} \quad (44)$$

$${}_s\boldsymbol{r}^{(i)}_t=\begin{bmatrix}{}_sr^{(i)}_t(1)\\ \vdots\\ {}_sr^{(i)}_t(P)\end{bmatrix} \quad (45)$$

where

$${}_sr^{(i)}_t(k)=\frac{1}{L}\sum_{l=0}^{L-1}V^{(i)}_{t,l}e^{j\frac{2\pi l}{N}k} \quad (46)$$
$$V^{(i)}_{t,l}=\left[1\ -{}_l\hat{\boldsymbol{g}}^{(i)H}\right]\left\langle\boldsymbol{X}_{t:t-K_l,l}\boldsymbol{X}^H_{t:t-K_l,l}\right\rangle_{p\left(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta\right)}\times\begin{bmatrix}1\\ -{}_l\hat{\boldsymbol{g}}^{(i)}\end{bmatrix} \quad (47)$$
$$\boldsymbol{X}_{t:t-k,l}=[X_{t,l},\cdots,X_{t-k,l}]^{\top}. \quad (48)$$

Since $\boldsymbol{X}_{t:t-K_l,l}$ is a subvector of $\boldsymbol{X}_l$, and $\boldsymbol{X}_l$ is Gaussianly distributed as shown in (40), the expectation in (47) can be calculated as

$$\left\langle\boldsymbol{X}_{t:t-K_l,l}\boldsymbol{X}^H_{t:t-K_l,l}\right\rangle_{p\left(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta\right)}=\boldsymbol{\mu}^{(i)}_{t:t-K_l,l}\boldsymbol{\mu}^{(i)}_{t:t-K_l,l}{}^{H}+\Sigma^{(i)}_{(t:t-K_l,t:t-K_l),l}. \quad (49)$$

The derivation of (42) and (43) will be described in Appendix C.

The update rules (42) and (43) may be interpreted as follows. We can find that ${}_sr^{(i)}_t(k)$ in (46) is an expected kth correlation coefficient of a time-domain clean speech signal in the tth frame. Hence, (42) can be considered as performing linear predictive analysis (LPA) by using the expected speech correlation coefficients. It can also be found that (43) calculates the corresponding prediction residual. If a clean speech signal were available, its all-pole parameters could be calculated with the LPA from the correlation coefficients of the clean speech signal. However, since the clean speech signal is inaccessible in reality, here the expected correlation coefficients are substituted for the true correlation coefficients.

D. CM-step2

Finally, we describe CM-step2, namely the update rule for estimates of room regression coefficients ${}_g\Theta$. Let us put all the room regression coefficients of the lth frequency band in a vector as

$$\boldsymbol{g}_l=[g_{1,l},\cdots,g_{K_l,l}]^{\top}. \quad (50)$$

Note that ${}_g\Theta$ coincides with the set of $\boldsymbol{g}_l$ over all the frequency bands.

Maximization step (39) is accomplished by updating the estimate of $\boldsymbol{g}_l$, for all frequency-band indices l, according to the following rule:

$$\hat{\boldsymbol{g}}^{(i+1)}_l={}_xR^{(i)}_l{}^{-1}\,{}_x\boldsymbol{r}^{(i)}_l. \quad (51)$$

${}_xR^{(i)}_l$ and ${}_x\boldsymbol{r}^{(i)}_l$ in (51) are defined as

$${}_xR^{(i)}_l=\sum_{t=0}^{T-1}\frac{A_t\left(e^{j\frac{2\pi l}{L}};\hat{\boldsymbol{a}}^{(i+1)}_t\right)}{{}_s\hat{\sigma}^{2\,(i+1)}_t}\times\left\langle\boldsymbol{X}_{t-1:t-K_l,l}\boldsymbol{X}^H_{t-1:t-K_l,l}\right\rangle_{p(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta)} \quad (52)$$

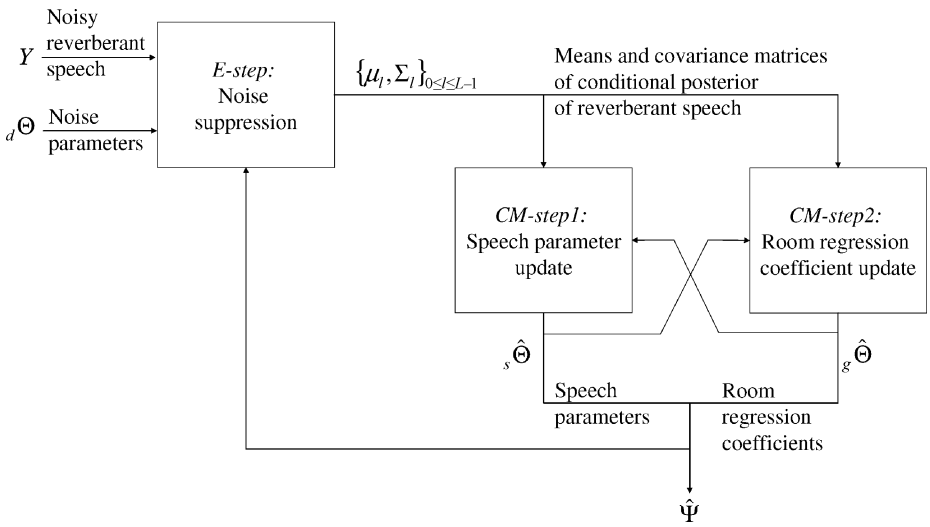

Fig. 4. Schematic diagram of parameter estimation process.

$$ {}_x\boldsymbol{r}_l^{(i)} = \sum_{t=0}^{T-1} \frac{A_t\left(e^{j\frac{2\pi l}{L}}; \hat{\boldsymbol{a}}_t^{(i+1)}\right)}{{}_s\hat{\sigma}_t^{2\,(i+1)}} \times \left\langle \boldsymbol{X}_{t-1:t-K_l,l} X_{t,l}^* \right\rangle_{p\left(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta\right)} \quad (53) $$

where $\boldsymbol{X}_{t-1:t-K_l,l}$ is defined in analogy to (48). The expectation in (52) and (53) is calculated as

$$ \left\langle \boldsymbol{X}_{t-1:t-K_l,l} \boldsymbol{X}_{t-1:t-K_l,l}^H \right\rangle_{p\left(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta\right)} = \boldsymbol{\mu}_{t-1:t-K_l,l}^{(i)} \boldsymbol{\mu}_{t-1:t-K_l,l}^{(i)\,H} + \Sigma_{(t-1:t-K_l,t-1:t-K_l),l}^{(i)} \quad (54) $$

$$ \left\langle \boldsymbol{X}_{t-1:t-K_l,l} X_{t,l}^* \right\rangle_{p\left(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta\right)} = \boldsymbol{\mu}_{t-1:t-K_l,l}^{(i)} \mu_{t,l}^{(i)\,*} + \Sigma_{(t-1:t-K_l,t:t),l}^{(i)}. \quad (55) $$

See Appendix D for the derivation of (51). ${}_xR_l^{(i)}$ and ${}_x\boldsymbol{r}_l^{(i)}$ can be computed efficiently by using a fast Fourier transform (FFT), although we omit the complicated details.

The update rule (51) can be interpreted as follows. ${}_xR_l^{(i)}$, defined by (52), is the expected value of the weighted covariance matrix of noise-free reverberant speech spectral component $X_{t,l}$ with lags of $0, \cdots, K_l - 1$. On the other hand, ${}_x\boldsymbol{r}_l^{(i)}$, defined by (53), is the expected value of the weighted covariance vector of $X_{t,l}$ with lags of $1, \cdots K_l$. Therefore, (51) may be viewed as modified Yule–Walker equations. The important point regarding this modification is that the weights are equal to the reciprocals of the speech PSD estimates. Such weighting is also involved in ML-based speech dereverberation methods [17]–[20], and was proven to be vital for speech dereverberation [19]. Because the noise-free reverberant speech spectral components are unavailable here, the expected weighted covariances are used.

E. Interpretation

Fig. 4 shows a schematic diagram of the ECM algorithm based parameter estimation process. As explained above (and as shown in Fig. 4), E-step suppresses noise components in observed noisy reverberant speech to yield the frequency bandwise means and covariance matrices of noise-free reverberant speech. On the other hand, the two CM-steps update the estimates of speech parameters and room regression coefficients. Existing dereverberation methods such as [16] and [19] are also composed of update processes for the speech parameters and room regression coefficients. In this sense, the set of the two CM-steps is responsible for dereverberation. Therefore, the proposed parameter estimation algorithm can be interpreted as alternately performing noise suppression and dereverberation processes.

As regards the proposed parameter estimation algorithm, we emphasize the following two points.

- The noise suppression process (E-step) and the dereverberation process (CM-step1 and CM-step2) are performed so that the likelihood function value increases monotonically. In other words, the noise suppression and dereverberation processes are integrated in the proposed algorithm under a single objective function, i.e., the likelihood function.
- The noise suppression process (E-step) outputs not only the means of the conditional posterior of the reverberant speech spectral components but also their covariance matrices. As will be shown in Section VII, the covariances must be taken into account if we are to improve the speech enhancement performance.

VI. Modifications and Summary of Proposed Speech Enhancement Method

Now, we reach a point where the above components can be assembled to derive the entire algorithm of the proposed method. Before we present the proposed algorithm, however, we introduce two modifications: one in the time–frequency analysis (and synthesis), and the other in the calculation of the conditional posterior, $p(\mathcal{X}|\mathcal{Y};\Theta)$, of reverberant speech spectral components.

A. Modification of Time–Frequency Analysis

The first modification is concerned with the time–frequency analysis. In Section II, we stated that the time–frequency analysis is performed by using STFT with an L-point frame length and a W-point frame shift. By capitalizing on the strong correspondence between the STFT representation and polyphase filter bank (PFB) [36], it would be acceptable to substitute

the PFB analysis as a time–frequency analysis method. Frame length L and frame shift W correspond to the number of filter banks and the decimation factor, respectively (note that the decimation factor is usually defined not as W but as L/W). For the purpose of ironing out differences over terminology, we call L and W frame length and frame shift even in the context of PFB. When PFB analysis is used for time–frequency analysis, time-domain signal synthesis is realized by PFB synthesis instead of overlap-add synthesis.

The merits and demerits of this substitution can be summarized as follows.

- Our experience shows that the PFB allows us to choose a larger value for the frame shift W than the STFT representation as long as an appropriate prototype low-pass filter is used. Indeed, our informal experiments revealed that the STFT representation required oversampling by a factor of at least four to achieve high-performance dereverberation in the noiseless case. By contrast, when the PFB was used, oversampling by a factor of two sufficed to obtain the same level of dereverberation performance.[2] The reason may be that the transition bandwidth of the PFB analysis can be made narrower than that of the STFT. An increase in the frame shift not only reduces the number of frames T but also enables us to choose a smaller value for regression order K_l. As a consequence, the use of the PFB makes it possible to reduce computational cost.
- One theoretical defect is that the concept of the PFB may disagree with the concept of power spectral density, which forms the basis for the speech and noise models. Nonetheless, in practice, this theoretical defect does not appear to degrade the performance as indicated by the experimental results in Section VII.

B. Modification of Conditional Posterior Calculation

The second modification is concerned with the calculation of the conditional posterior of reverberant speech spectral components, $p(\mathcal{X}|\mathcal{Y};\Theta)$. This calculation appears in E-step and the MMSE clean speech spectrum estimation. Mean $\boldsymbol{\mu}_l$ and covariance matrix Σ_l of $p(\mathcal{X}|\mathcal{Y};\Theta)$ are theoretically defined as (25) and (26). However, we experimentally found that calculating $\boldsymbol{\mu}_l$ and Σ_l in exact accordance with (25) and (26) often made matrix $(B_l B_l^H + G_l A_l A_l^H G_l^H)$ nearly rank deficient. This phenomenon results in the divergence of the parameter estimates.

In order to stabilize the behavior of the proposed algorithm, we modify the formulas for calculating $\boldsymbol{\mu}_l$ and Σ_l as follows. Let $\rho_{m,n}$ be the (m,n)th element of $(B_l B_l^H + G_l A_l A_l^H G_l^H)$. We propose replacing $(B_l B_l^H + G_l A_l A_l^H G_l^H)$ by diagonal matrix $\mathrm{diag}\{\rho_{1,1},\cdots,\rho_{T,T}\}$. Therefore, (25) and (26) are replaced respectively by

$$\boldsymbol{\mu}_l = \mathrm{diag}\left\{\frac{1}{\rho_{1,1}},\cdots,\frac{1}{\rho_{T,T}}\right\}\left(B_l B_l^H\right)\boldsymbol{Y}_l \tag{56}$$

$$\Sigma_l = \mathrm{diag}\left\{\frac{1}{\rho_{1,1}},\cdots,\frac{1}{\rho_{T,T}}\right\}. \tag{57}$$

[2]The difference in performance between the STFT representation and the PFB became small as the magnitude of the additive noise increased. Nonetheless, the STFT representation never outperformed the PFB significantly. Thus, we recommend oversampling by a factor of four when the STFT representation is used, and by a factor of two with the PFB.

Although this modification may detract from the monotonic increase of the likelihood function value, we did not observe any disruptive effects (see Section VII).

This modification enjoys another advantage; forcing $(B_l B_l^H + G_l A_l A_l^H G_l^H)$ to be diagonal contributes to the reduction in computational cost. This is because the inversion of a T-dimensional square matrix in (25) and (26) is avoided. For these two reasons, we substitute (56) and (57) for (25) and (26), respectively.

C. Summary of Proposed Algorithm

Based on all the above derivations, the proposed algorithm is summarized as follows.

Proposed speech enhancement algorithm

1) Analysis
   Observed noisy reverberant speech signal $y(n)$ is transformed into the time–frequency domain by using STFT or PFB analysis.
2) Parameter estimation
   a) Initialize the estimates of unknown parameters Ψ as $\hat{\Psi}^{(0)} = \{{}_s\hat{\Theta}^{(0)}, {}_g\hat{\Theta}^{(0)}\}$, and set iteration index i at 0.
   b) Calculate $\boldsymbol{\mu}_l(\hat{\Psi}^{(i)}, {}_d\Theta, \mathcal{Y})$ and $\Sigma_l(\hat{\Psi}^{(i)}, {}_d\Theta)$ according to (56) and (57), respectively, for all frequency-band indices l.
   c) Update the estimates of speech parameters ${}_s\Theta$ to ${}_s\hat{\Theta}^{(i+1)}$ by calculating (42) and (43) for all frame indices t.
   d) Update the estimates of channel regression coefficients ${}_g\Theta$ to ${}_g\hat{\Theta}^{(i+1)}$ by calculating (51) for all frequency-band indices l. The parameter estimates after the $(i+1)$th iteration are obtained as $\hat{\Psi}^{(i+1)} = \{{}_s\hat{\Theta}^{(i+1)}, {}_s\hat{\Theta}^{(i+1)}\}$.
   e) Increment i by 1, and then return to b) unless convergence is reached.
   f) Set the final parameter estimates as $\hat{\Psi} = \hat{\Psi}^{(i)}$.
3) Clean speech spectrum estimation
   Calculate MMSE estimate $\hat{S}_{t,l}$ of clean speech spectral component $S_{t,l}$ according to (34) by using the estimated parameters, $\hat{\Psi}$, and the known noise parameters, ${}_d\Theta$.
4) Synthesis
   Finally, the estimate of the clean speech spectral component, $\hat{S}_{t,l}$, is transformed into the time domain to synthesize the estimate of clean speech signal $s(n)$. The transformation is accomplished by using overlap-add synthesis for STFT or PFB synthesis for PFB analysis.

VII. Experimental Results

We conducted experiments to evaluate the performance of the proposed method. Sections VII-A to VII-D closely examine the fundamental performance and several aspects of the proposed method based on experimental results obtained using one reverberation condition and white noise. Furthermore, Section VII-E reports experimental results obtained using various reverberation conditions and pink noise.

A. Experimental Task

We took Japanese utterances spoken by ten speakers from the ASJ-JNAS database [37]. There were five male and five female speakers. Fifty utterances were used for each speaker. Hence, a total of 500 utterances were used whose lengths ranged from 1.75 to 12.23 s. The sampling rate was 8 kHz. The acoustic signals of individual utterances were convolved with a room impulse response measured in a room with an RT_{60} of 0.6 s to synthesize the corresponding noise-free reverberant signals. The reverberation time index RT_{60} is the time needed for the sound pressure level to decay by 60 dB. Then, these reverberant signals were superimposed by a white Gaussian noise signal synthesized on a computer. The noise magnitude was adjusted so that the reverberant signal-to-noise ratio (RSNR) become 20, 15, or 10 dB. The RSNR is specifically defined as

$$\mathrm{RSNR} = 10\log_{10}\frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1} d(n)^2} \quad (\mathrm{dB}) \tag{58}$$

where N is the number of samples, and $x(n)$ and $d(n)$ are a noise-free reverberant speech signal and a noise signal, respectively. Thus, 1500 noisy reverberant speech signals were generated in total.

The experimental task was to estimate the corresponding clean speech signal in an offline manner when one of the 1500 noisy reverberant speech signals was presented. Therefore, the experiment involved a total of 1500 trials. The results of individual trials were evaluated in terms of Mel-frequency cepstral coefficient (MFCC) distance. The MFCC distance was chosen since it is related to both the audible quality of speech and the automatic speech recognition performance. Applying the proposed method to an automatic speech recognition task and evaluating the results are subjects for future work. The MFCC distance is calculated according to the following procedure.

1) Divide clean speech signal $s(n)$ and its estimate $\hat{s}(n)$ into short-time segments by using a 256-point Hann window with a 128-point overlap. Note that here the frame size and frame shift may be different from those of a speech enhancement system.
2) Calculate 12th-order MFCCs (including zeroth order components) of individual segmented signals. The respective MFCC sequences of $s(n)$ and $\hat{s}(n)$ were further processed with cepstral mean normalization (CMS). CMS was used to exclude from the evaluation the distortion caused by the room's early reflections, which cannot be cancelled out by the proposed method as pointed out in Section III-A.
3) Calculate MFCC distance D_{MFCC} by

$$D_{\mathrm{MFCC}} = \frac{1}{T'}\sum_{t=0}^{T'-1}\sum_{k=0}^{12}(c_{t,l} - \hat{c}_{t,k})^2 \tag{59}$$

where T' is the number of short-time segments, and $c_{t,k}$ and $\hat{c}_{t,k}$ are the kth mean-normalized MFCCs of $s(n)$ and $\hat{s}(n)$, respectively, at the tth frame.

B. System Setup

A speech enhancement system based on the proposed method was implemented as a MATLAB program. The program was run on a Linux PC equipped with a 3.6-GHz Pentium 4 processor.

TABLE I
PARAMETER SETTINGS

Analysis	–	STFT or PFB analysis
Frame size	L	256
Frame shift	W	64 (STFT) or 128 (PFB)
Number of poles	P	12
Regression order	K_l	50 (STFT) or 30 (PFB)
Number of ECM iterations	–	5

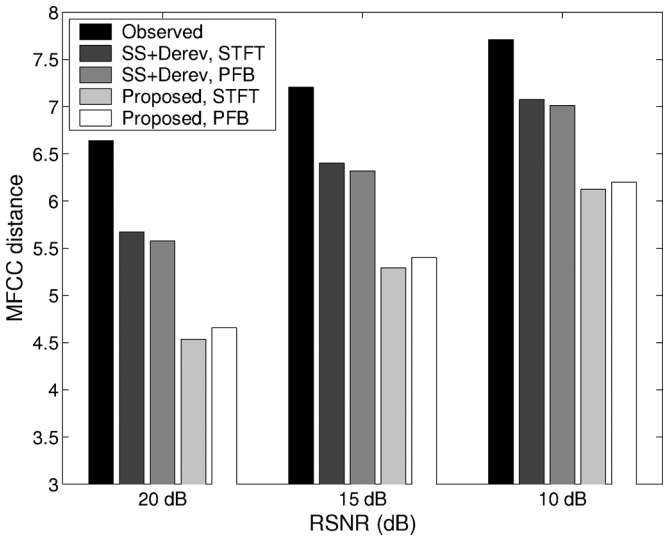

Fig. 5. MFCC distances averaged over utterances.

As regards the time–frequency analysis, both STFT and PFB analysis were tested. A Hann window was used for the STFT while a $(2L+1)$-tap prototype filter was used for the PFB. Frame size L, frame shift W, and regression order K_l were set at 256, 64, and 50, respectively, when STFT was used.[3] On the other hand, with PFB, L, W, and K_l were set at 256, 128, and 30, respectively. The number of poles P was 12, and there were five iterations for the ECM algorithm. These settings are listed in Table I.

The initial parameter estimates $\hat{\Psi}^{(0)} = \{{}_s\hat{\Theta}^{(0)}, {}_g\hat{\Theta}^{(0)}\}$ were given as follows. Observed noisy reverberant speech spectra were first processed with spectral subtraction [24]. Then, for each frame index t, initial speech parameter estimates $\hat{\boldsymbol{a}}_t^{(0)}$ and ${}_s\hat{\sigma}_t^{2(0)}$, respectively, were set at the LPCs and prediction residual calculated from the noise-suppressed reverberant speech spectrum at the tth frame. ${}_s\Theta^{(0)}$ was obtained as ${}_s\Theta^{(0)} = \{\hat{\boldsymbol{a}}_t^{(0)}, {}_s\hat{\sigma}_t^{2(0)}\}_{0\le t\le T-1}$. Finally, for each frequency-band index l, initial estimates $\hat{\boldsymbol{g}}_l^{(0)}$ of room regression coefficients were calculated by performing CM-step2 (51) once, where $\boldsymbol{\mu}_{T-1:0,l}^{(-1)}$ and $\Sigma_{(T-1:0,T-1:0),l}^{(-1)}$ were set at the vector of the noise-suppressed reverberant speech spectral components and a zero matrix, respectively. ${}_g\Theta^{(0)}$ was obtained as ${}_g\Theta^{(0)} = \{\hat{\boldsymbol{g}}_l^{(0)}\}_{0\le l\le L-1}$.

As a competing method, we also implemented a speech enhancement system that sequentially performs noise suppression and dereverberation processes. The noise suppression was accomplished with spectral subtraction. Spectral subtraction

[3]Setting the regression order K_l at larger values did not improve the speech enhancement performance.

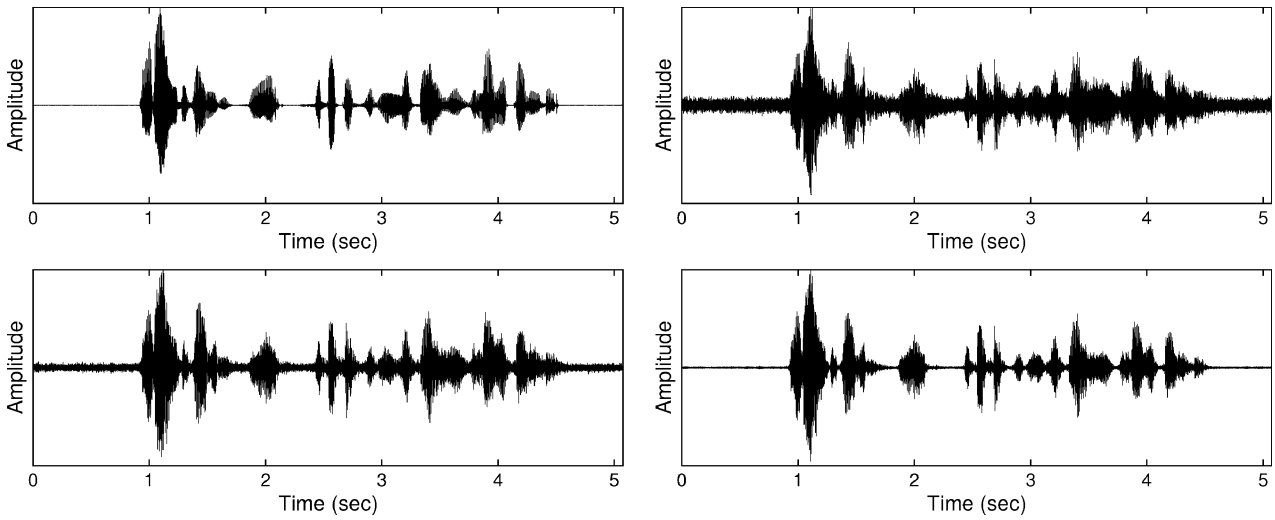

Fig. 6. Waveforms of clean speech (top left), noisy reverberant speech (top right), speech estimated with the competing method (bottom left), and speech estimated with the proposed method (bottom right). The RSNR is 10 dB.

was chosen because it is used by existing noisy reverberant speech enhancement method [22]. For the dereverberation, we used the frequency-domain dereverberation method proposed in [20], which was slightly modified so that the short-time PSD of speech is represented by the all-pole model.

The proposed and competing methods have the following differences.

a1) The proposed method iteratively updates estimates of noise-free reverberant speech spectral components based on estimates of speech parameters and channel regression coefficients, which are also iteratively updated in the ECM algorithm. By contrast, the competing method calculates estimates of reverberant spectral components only once.

a2) The proposed method calculates estimates of reverberant speech spectral components in the form of a probability distribution [see (40)], while the competing method performs a point estimation of the reverberant speech spectral components.

Other features are common to both methods. Hence, a comparison of the results obtained with both methods will highlight the effects of the above two differences.

C. Results

Fig. 5 shows the MFCC distances averaged over the 500 utterances for individual system settings. The results are grouped in terms of RSNR. It is clear that the proposed method substantially outperformed the competing method for all the RSNR conditions. There was little difference between the performance with the STFT and that with the PFB analysis.

In Fig. 6, we plot example waveforms of clean speech, noisy reverberant speech, the output speech of the competing method, and the output speech of the proposed method. These waveforms are for the PFB analysis and the 10-dB RSNR condition. The MFCC distances of the noisy reverberant speech, the competing method's output, and the proposed method's output were 7.30, 6.49, and 5.30, respectively. It is clear that the proposed method accentuated speech components more distinctly than the competing method. Audio examples are available at the author's website.[4]

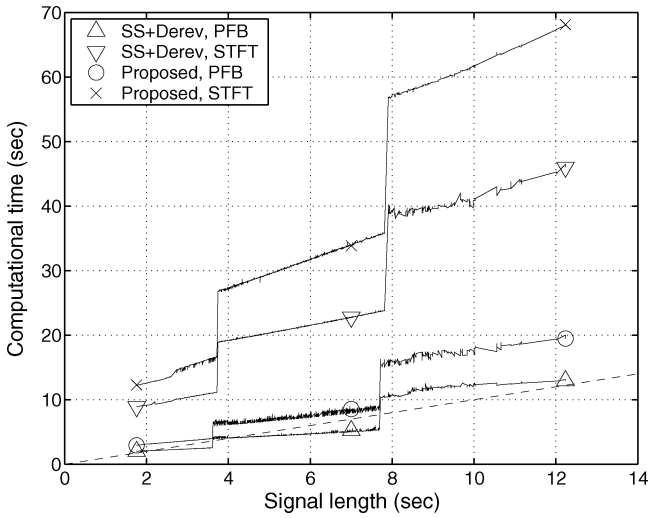

Fig. 7. Computational time versus signal length curve. The dashed line shows the boundary where computational time matches observed signal length.

Next, we investigate the computational cost of the proposed method. Fig. 7 shows the computational time as a function of the observed signal length. It can be clearly seen that the computational time with the PFB analysis was much faster than that with the STFT. In addition, when the PFB analysis was used, the extra computational time required by the proposed method was sufficiently small.

From the above results, we conclude that the proposed method with the PFB analysis is the best choice as regards enhancing noisy reverberant speech.

[4][Online]. Available: http://www.kecl.ntt.co.jp/icl/signal/takuya/research.html

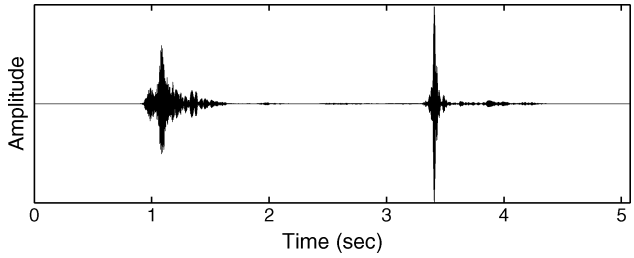

Fig. 8. Speech waveform estimated by forcing covariance matrix Σ_l to be a zero matrix.

D. Discussion

In this section, we discuss certain aspects of the proposed method. The following discussion is based on the experimental results obtained with the PFB.

Let us consider forcing the covariance matrix, Σ_l in (57), of the conditional posterior of reverberant speech spectral components to be a zero matrix. Then, the average MFCC distances degraded to 14.8, 15.5, and 17.3 for RSNRs of 20, 15, and 10 dB, respectively. Fig. 8 shows the estimated speech waveform for the same observed signal as that in Fig. 6. We can clearly see that the noise components were excessively suppressed and that the speech enhancement was not achieved at all. We found that, in general, the degradation of speech quality became severe with the increase in the number of iterations. Similar phenomena were observed in the all-pole model based noise suppression [2]. This result indicates that we must take the covariances of the conditional posterior of reverberant speech into account in the proposed method. In other words, both characteristics of the proposed method described in Section VII-B, a1) and a2), combine to improve the speech enhancement performance.

Fig. 9 plots the MFCC distance improvements for individual genders grouped in terms of RSNR. It can be seen that although the MFCC distance improvements for male speech were somewhat greater than those for female speech, there was little difference between them. The all-pole model is known to be less suitable for female speech than male speech because female speech generally has higher fundamental frequencies. However, from Fig. 9, it can be concluded that the proposed method is effective regardless of the speaker's gender.

Fig. 10 shows the MFCC distance improvements obtained in respective trials plotted against the observed signal length. We can find that the proposed method never increased the MFCC distances. From Fig. 10, we can also see the relationship between the observed signal length and the performance of the proposed method. The MFCC distance improvements tended to increase moderately as the observed signals became longer regardless of RSNR. Nevertheless, the proposed method worked well even when the observed signals were shorter than 5 s.

E. Experimental Results Using Other Conditions

We further tested the proposed method by using other reverberation and noise conditions. For clean speech signals, we used the same signals as in the preceding sections. Ten room impulse responses were used. These impulse responses were different from each other as regards speaker-to-microphone distances and RT_{60}s. The speaker-to-microphone distance was 1 or 2 m, and

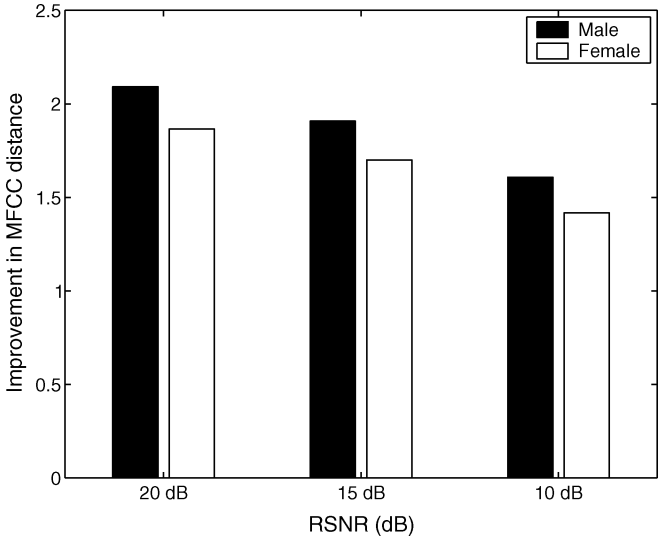

Fig. 9. MFCC distance improvements averaged over utterances shown by gender.

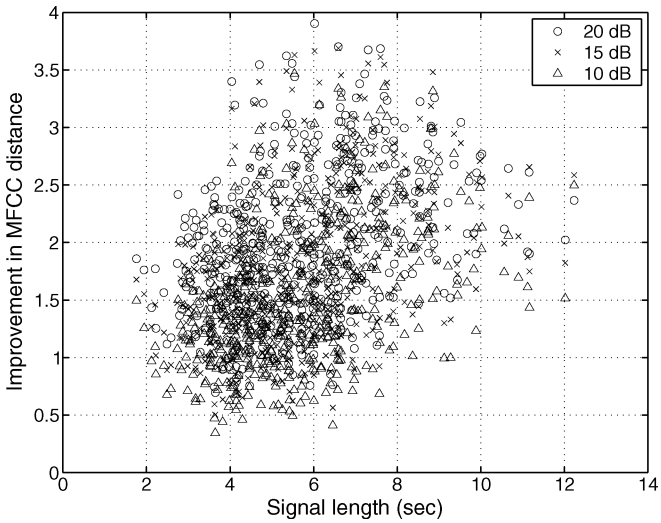

Fig. 10. Distribution of individual results for observed signal length versus MFCC distance improvement plane.

the RT_{60} was 0.15, 0.2, 0.3, 0.4, and 0.5 s. We used white noise and pink noise as noise signals. The RSNR was fixed at 15 dB.

Fig. 11 shows the average MFCC distances as a function of RT_{60} when the white noise was used. It can be seen that the improvements in the MFCC distances were comparable for all RT_{60} values. Fig. 12 shows the average MFCC distances as a function of RT_{60} with the pink noise. Again, the MFCC distances were improved for all RT_{60} values. These results demonstrate the effectiveness of the proposed method under various reverberation and noise conditions.

VIII. Conclusion

This paper described the problem of integrating noise suppression and dereverberation processes, and proposed an enhancement method for noisy reverberant speech signals. The significant characteristics of the proposed method are as follows.

- Noise suppression and dereverberation processes are performed alternately.

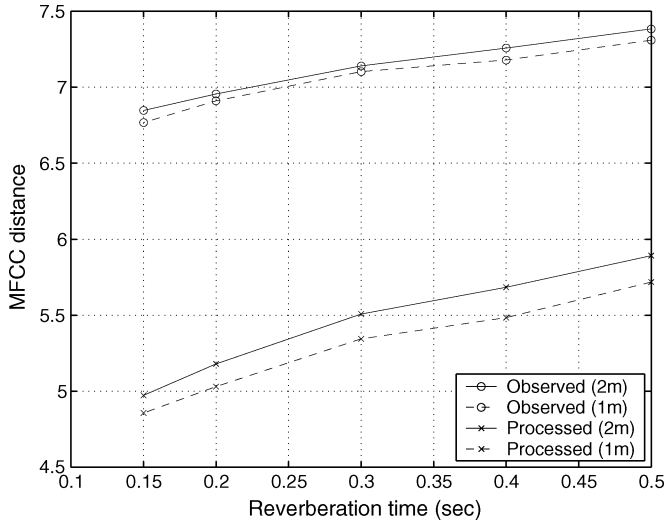

Fig. 11. MFCC distances as a function of RT_{60} when using white noise.

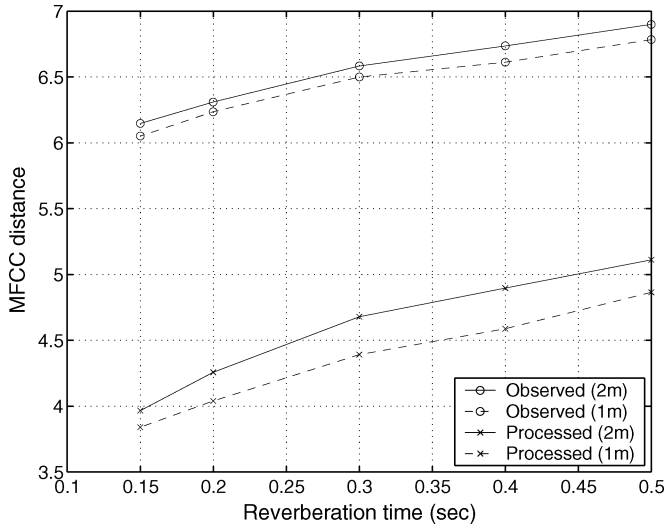

Fig. 12. MFCC distances as a function of RT_{60} when using pink noise.

- Noise-free reverberant speech spectra are estimated in the form of a probability distribution. This means that the proposed method takes account of the reliability of the estimates of the noise-free reverberant speech spectra.

These characteristics are the direct outcome of the fact that the proposed method is based on ML estimation. The following two factors in the observed noisy reverberant speech modeling enabled us to derive an effective and efficient algorithm.

- As the speech model, we used an all-pole model, which has been successfully employed for both noise suppression and dereverberation.
- We modeled the room's convolutive system as a frequency band-wise AR system, which allowed the easy integration of the noise suppression and dereverberation processes.

Experimental results revealed the benefit of the above characteristics. Indeed, the proposed method outperformed a method in which noise suppression and dereverberation systems are connected in tandem.

Future work includes the following issues.

- *Adaptive estimation of room regression coefficients*: In a practical situation, the convolutive system of a room's acoustic system often varies with time due, for example, to changes in speaker positions. Hence, the estimates of the room regression coefficients must be updated adaptively.
- *Adaptive estimation of noise PSD*: Noise PSD may also change with time in a practical situation. Therefore, the adaptive estimation of the noise PSD is also a significant research challenge.

We believe that we can also address these issues based on a statistical approach as with the proposed method.

APPENDIX

Here, we describe the derivations omitted from the body of this paper.

A. Derivation of Joint PDF of Noisy and Noise-Free Reverberant Speech Spectral Components

In this section, we show that the joint pdf of noisy reverberant speech spectral components $\mathcal{Y}$ and noise-free reverberant speech spectral components $\mathcal{X}$, $p(\mathcal{Y},\mathcal{X};\Theta)$, is given by (14).

$p(\mathcal{Y},\mathcal{X};\Theta)$ is factorized as

$$p(\mathcal{Y},\mathcal{X};\Theta) = p(\mathcal{Y}|\mathcal{X};{}_d\Theta)p(\mathcal{X};\Psi). \tag{60}$$

Based on noise model (13), the first term on the right-hand side of (60) is given by

$$\begin{aligned} p(\mathcal{Y}|\mathcal{X};{}_d\Theta) &= \prod_{t=0}^{T-1}\prod_{l=0}^{L-1}\mathcal{N}_{\mathbb{C}}\left\{Y_{t,l};X_{t,l},{}_d\lambda\left(\frac{2\pi l}{L}\right)\right\} \\ &= \prod_{t=0}^{T-1}\prod_{l=0}^{L-1}\frac{1}{\pi_d\lambda(2\pi l/L)} \\ &\quad\times\exp\left\{-\frac{|Y_{t,l}-X_{t,l}|^2}{{}_d\lambda(2\pi l/L)}\right\}. \end{aligned} \tag{61}$$

On the other hand, the second term on the right hand side of (60) is written as

$$\begin{aligned} p(\mathcal{X};\Psi) &= \prod_{t=0}^{T-1}\prod_{l=0}^{L-1}p\left(X_{t,l}|X_{t-1,l},\cdots,X_{t-K_l,l};\Psi\right) \\ &= \prod_{t=0}^{T-1}\prod_{l=0}^{L-1}\mathcal{N}_{\mathbb{C}}\left\{X_{t,l};\sum_{k=1}^{K_l}g_{k,l}^{*}X_{t-k,l},{}_s\lambda_t\left(\frac{2\pi l}{L}\right)\right\} \\ &= \prod_{t=0}^{T-1}\prod_{l=0}^{L-1}\frac{1}{\pi}\frac{\left|A_t\left(e^{j\frac{2\pi l}{L}}\right)\right|^2}{{}_s\sigma_t^2} \\ &\quad\times\exp\left\{-\frac{\left|A_t\left(e^{j\frac{2\pi l}{L}}\right)\right|^2}{{}_s\sigma_t^2}\right. \\ &\qquad\left.\times\left|X_{t,l}-\sum_{k=1}^{K_l}g_{k,l}^{*}X_{t-k,l}\right|^2\right\}. \end{aligned} \tag{62}$$

The second line of (62) is obtained based on reverberation model (7) and speech models (9) and (10). By substituting (61) and (62) into (60), we obtain (14).

B. Derivation of Conditional Posterior of Reverberant Speech Spectral Components

In this section, we show that the conditional posterior of reverberant speech spectral components, $p(\mathcal{X}|\mathcal{Y};\Theta)$, is given by (24).

The conditional posterior $p(\mathcal{X}|\mathcal{Y};\Theta)$ is proportional to $p(\mathcal{Y},\mathcal{X};\Theta)$ because of the following equation:

$$p(\mathcal{X}|\mathcal{Y};\Theta) = \frac{p(\mathcal{Y},\mathcal{X};\Theta)}{p(\mathcal{Y};\Theta)} \propto p(\mathcal{Y},\mathcal{X};\Theta) \quad \text{with respect to } \mathcal{X}. \tag{63}$$

Hence, we reorganize the joint pdf, $p(\mathcal{Y},\mathcal{X};\Theta)$, of noisy and noise-free reverberant speech spectral components with respect to $\mathcal{X}$.

The joint pdf $p(\mathcal{Y},\mathcal{X};\Theta)$ is given by (20). The term within the braces of (20) is reorganized as

$$\begin{aligned}
&\sum_{t=0}^{T-1}\left(\frac{|Y_{t,l}-X_{t,l}|^2}{{}_d\lambda(2\pi l/L)}+\frac{\left|A_t\left(e^{j\frac{2\pi l}{L}}\right)\right|^2}{{}_s\sigma_t^2}\right.\\
&\qquad\left.\times\left|X_{t,l}-\sum_{k=1}^{K_l} g_{k,l}^{*}X_{t-k,l}\right|^2\right)\\
&=(\boldsymbol{Y}_l-\boldsymbol{X}_l)^H B_l B_l^H(\boldsymbol{Y}_l-\boldsymbol{X}_l)+\boldsymbol{X}_l^H G_l A_l A_l^H G_l^H \boldsymbol{X}_l\\
&=\left\{\boldsymbol{X}_l-\left(B_lB_l^H+G_lA_lA_l^HG_l^H\right)^{-1}\left(B_lB_l^H\right)\boldsymbol{Y}_l\right\}\\
&\quad\times\left(B_lB_l^H+G_lA_lA_l^HG_l^H\right)\\
&\quad\times\left\{\boldsymbol{X}_l-\left(B_lB_l^H+G_lA_lA_l^HG_l^H\right)^{-1}\left(B_lB_l^H\right)\boldsymbol{Y}_l\right\}\\
&\quad+\text{constant (with respect to } \boldsymbol{X}_l),
\end{aligned} \tag{64}$$

where $\boldsymbol{X}_l, \boldsymbol{Y}_l, G_l, A_l$, and B_l are given by (22), (23), (27), (28), and (29), respectively. To get from the third line to the fourth line, we used the fact that $B_lB_l^H+G_lA_lA_l^HG_l^H$ is an Hermitian matrix. By using (63) and (64), we obtain

$$\begin{aligned}
&p(\mathcal{X}|\mathcal{Y};\Theta)\\
&\propto\prod_{l=0}^{L-1}\exp\left\{\left\{\boldsymbol{X}_l-\left(B_lB_l^H+G_lA_lA_l^HG_l^H\right)^{-1}\left(B_lB_l^H\right)\boldsymbol{Y}_l\right\}\right.\\
&\qquad\times\left(B_lB_l^H+G_lA_lA_l^HG_l^H\right)\\
&\qquad\left.\times\left\{\boldsymbol{X}_l-\left(B_lB_l^H+G_lA_lA_l^HG_l^H\right)^{-1}\right.\right.\\
&\qquad\left.\left.\times\left(B_lB_l^H\right)\boldsymbol{Y}_l\right\}\right\}.
\end{aligned} \tag{65}$$

This equation indicates that the conditional posterior $p(\mathcal{X}|\mathcal{Y};\Theta)$ is given by (24).

C. Derivation of CM-step1

In this section, we provide a solution for (38) to derive the update rules of CM-step1, (42) and (43). Below, for brevity, we sometimes omit the description of pdf $p(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta)$, which appears in the definition of auxiliary function $Q(\Psi;\hat{\Psi}^{(i)})$.

By substituting (20) into (36), we can write the auxiliary function as

$$\begin{aligned}
&Q\left(\Psi;\hat{\Psi}^{(i)}\right)\\
&=\left\langle\log p(\mathcal{Y},\mathcal{X};\Psi,{}_d\Theta)\right\rangle\\
&=-\sum_{t=0}^{T-1}\left(L\log\left({}_s\sigma_t^2\right)+\frac{1}{{}_s\sigma_t^2}\sum_{l=0}^{L-1}\left|A_t\left(e^{j\frac{2\pi l}{L}};\boldsymbol{a}_t\right)\right|^2\right.\\
&\qquad\left.\times\left\langle\left|X_{t,l}-\sum_{k=1}^{K_l}g_{k,l}^{*}X_{t-k,l}\right|^2\right\rangle\right).
\end{aligned} \tag{66}$$

The fourth line of (66) is rewritten as

$$\begin{aligned}
&\left\langle\left|X_{t,l}-\sum_{k=1}^{K_l}g_{k,l}^{*}X_{t-k,l}\right|^2\right\rangle=\begin{bmatrix}1 & -\boldsymbol{g}_l^H\end{bmatrix}\\
&\qquad\times\left\langle\boldsymbol{X}_{t:t-K_l,l}\boldsymbol{X}_{t:t-K_l,l}^H\right\rangle\begin{bmatrix}1\\-\boldsymbol{g}_l\end{bmatrix}
\end{aligned} \tag{67}$$

where $\boldsymbol{X}_{t:t-k,l}$ is defined by (48). By substituting (67) and ${}_g\Theta={}_g\hat{\Theta}^{(i)}$ into (66), the function to be maximized is obtained as

$$\begin{aligned}
Q\left({}_s\Theta,{}_g\hat{\Theta}^{(i)};\hat{\Psi}^{(i)}\right)=&-\sum_{t=0}^{T-1}\left(L\log\left({}_s\sigma_t^2\right)\right.\\
&\left.+\frac{1}{{}_s\sigma_t^2}\sum_{l=0}^{L-1}\left|A_t\left(e^{j\frac{2\pi l}{L}};\boldsymbol{a}_t\right)\right|^2V_{t,l}^{(i)}\right)
\end{aligned} \tag{68}$$

where $V_{t,l}^{(i)}$ is defined by (47). It can be found that the term in the second line of (68) is the inverse Fourier transform of the product of $\left|A_t(e^{j(2\pi l/L)};\boldsymbol{a}_t)\right|^2$ and $V_{t,l}^{(i)}$. Thus, this term can be transformed as

$$\begin{aligned}
&\frac{1}{L}\sum_{l=0}^{L-1}\left|A_t\left(e^{j\frac{2\pi l}{L}};\boldsymbol{a}_t\right)\right|^2V_{t,l}^{(i)}\\
&=\begin{bmatrix}1 & -\boldsymbol{a}_t^{\top}\end{bmatrix}\begin{bmatrix}{}_sr_t^{(i)}(0) & {}_s\boldsymbol{r}_t^{(i)\top}\\ {}_s\boldsymbol{r}_t^{(i)} & {}_sR_t^{(i)}\end{bmatrix}\begin{bmatrix}1\\-\boldsymbol{a}_t\end{bmatrix}\\
&=\left(\boldsymbol{a}_t-{}_sR_t^{(i)-1}{}_s\boldsymbol{r}_t^{(i)}\right)^{\top}{}_sR_t^{(i)}\left(\boldsymbol{a}_t-{}_sR_t^{(i)-1}{}_s\boldsymbol{r}_t^{(i)}\right)\\
&\quad+{}_sr_t^{(i)}(0)-{}_s\boldsymbol{r}_t^{(i)\top}{}_sR_t^{(i)-1}{}_s\boldsymbol{r}_t^{(i)}
\end{aligned} \tag{69}$$

where ${}_sR_t^{(i)}$, ${}_s\boldsymbol{r}_t^{(i)}$, and ${}_sr_t^{(i)}(k)$ are defined by (44), (45), and (46), respectively. Plugging (69) into (68) yields

$$\begin{aligned}
&Q\left({}_s\Theta,{}_g\hat{\Theta}^{(i)};\hat{\Psi}^{(i)}\right)\\
&=-L\sum_{t=0}^{T-1}\left\{\log\left({}_s\sigma_t^2\right)+\frac{1}{{}_s\sigma_t^2}\right.\\
&\qquad\times\left(\left(\boldsymbol{a}_t-{}_sR_t^{(i)-1}{}_s\boldsymbol{r}_t^{(i)}\right)^{\top}{}_sR_t^{(i)}\right.\\
&\qquad\left.\left.\times\left(\boldsymbol{a}_t-{}_sR_t^{(i)-1}{}_s\boldsymbol{r}_t^{(i)}\right)+{}_sr_t^{(i)}(0)\right.\right.\\
&\qquad\left.\left.-{}_s\boldsymbol{r}_t^{(i)\top}{}_sR_t^{(i)-1}{}_s\boldsymbol{r}_t^{(i)}\right)\right\}.
\end{aligned} \tag{70}$$

Therefore, $\hat{\boldsymbol{a}}_t^{(i+1)}$ and ${}_s\hat{\sigma}_t^{2(i+1)}$ maximizing (70) can be obtained as (42) and (43), respectively.

D. Derivation of CM-step2

In this section, we provide a solution for (39) to derive the update rule of CM-step2, (51).

We begin with organizing auxiliary function $Q(\Psi;\hat{\Psi}^{(i)})$ with respect to channel regression parameters ${}_g\Theta$, which are variables in this context. Below, for brevity, we sometimes omit the description of pdf $p(\mathcal{X}|\mathcal{Y};\hat{\Psi}^{(i)},{}_d\Theta)$, which appears in the definition of the auxiliary function.

From (20) and (36), the auxiliary function can be written as

$$\begin{aligned}
&Q\left(\Psi;\hat{\Psi}^{(i)}\right)\\
&\quad=\left\langle\log p(\mathcal{Y},\mathcal{X};\Psi,{}_d\Theta)\right\rangle\\
&\quad=-\sum_{t=0}^{T-1}\sum_{l=0}^{L-1}\frac{\left|A_t\left(e^{j\frac{2\pi l}{L}};\boldsymbol{a}_t\right)\right|^2}{{}_s\sigma_t^2}\\
&\qquad\times\left\langle\left|X_{t,l}-\sum_{k=1}^{K_l}g_{k,l}^{*}X_{t-k,l}\right|^2\right\rangle\\
&\quad=-\sum_{t=0}^{T-1}\sum_{l=0}^{L-1}\frac{\left|A_t\left(e^{j\frac{2\pi l}{L}};\boldsymbol{a}_t\right)\right|^2}{{}_s\sigma_t^2}\\
&\qquad\times\left\langle\left(X_{t,l}-\boldsymbol{g}_l^H\boldsymbol{X}_{t-1:t-K_l,l}\right)\right.\\
&\qquad\qquad\left.\times\left(X_{t,l}-\boldsymbol{g}_l^H\boldsymbol{X}_{t-1:t-K_l,l}\right)^H\right\rangle\\
&\quad=-\sum_{t=0}^{T-1}\sum_{l=0}^{L-1}\frac{\left|A_t\left(e^{j\frac{2\pi l}{L}};\boldsymbol{a}_t\right)\right|^2}{{}_s\sigma_t^2}\\
&\qquad\times\left\langle\boldsymbol{g}_l^H\boldsymbol{X}_{t-1:t-K_l,l}\boldsymbol{X}_{t-1:t-K_l,l}^H\boldsymbol{g}_l-\boldsymbol{g}_l^H\boldsymbol{X}_{t-1:t-K_l,l}X_{t,l}^{*}\right.\\
&\qquad\qquad\left.-X_{t,l}^{*}\boldsymbol{X}_{t-1:t-K_l,l}\boldsymbol{g}_l\right\rangle\\
&\qquad+\text{constant (with respect to }{}_g\Theta)
\end{aligned}\tag{71}$$

where $\boldsymbol{X}_{t:t-k,l}$ is defined by (48). By letting ${}_s\Theta={}_s\hat{\Theta}^{(i+1)}$ in (71), the function to be maximized is finally obtained as

$$\begin{aligned}
Q\left({}_s\hat{\Theta}^{(i+1)},{}_g\Theta;\hat{\Psi}^{(i)}\right)=&-\sum_{l=0}^{L-1}\left(\boldsymbol{g}_l-{}_xR_l^{(i)^{-1}}{}_x\boldsymbol{r}_l^{(i)}\right)^H{}_xR_l\\
&\times\left(\boldsymbol{g}_l-{}_xR_l^{(i)^{-1}}{}_x\boldsymbol{r}_l^{(i)}\right)+\text{constant}
\end{aligned}\tag{72}$$

where ${}_xR_l^{(i)}$ and ${}_x\boldsymbol{r}_l^{(i)}$ are given by (52) and (53), respectively. Hence, it is obvious that $\hat{\boldsymbol{g}}_l^{(i+1)}$ that maximizes (72) is given by (51).

REFERENCES

[1] T. Quatieri, *Discrete-Time Speech Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 2002.

[2] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-26, no. 3, pp. 197–210, Jun. 1978.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 373–385, Jul. 1998.

[4] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[5] 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.

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

[7] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007, 10.1155/2007/34013, article ID 34013..

[8] S. Javidi, N. D. Gaubitch, and P. A. Naylor, "An experimental study of the eigendecomposition methods for blind SIMO system identification in the presence of noise," in *Proc. Int. Worksh. Acoust. Echo, Noise Contr.*, 2006, CD-ROM Proc..

[9] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[10] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. VI, pp. 3701–3704.

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

[12] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

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

[14] J. R. Hopgood and P. J. W. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 5, pp. 476–488, 2003.

[15] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[16] T. Yoshioka, T. Hikichi, and M. Miyoshi, "Dereverberation by using time-variant nature of speech production system," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007, 10.1155/2007/65698, article ID 65698.

[17] T. Yoshioka, T. Nakatani, T. Hikichi, and M. Miyoshi, "Overfitting-resistant speech dereverberation," in *Proc. IEEE Worksh. Appl. Signal Process. Audio, Acoust.*, 2007, pp. 163–166.

[18] T. Nakatani, B.-H. Juang, T. Hikichi, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Study on speech dereverberation with autocorrelation codebook," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, 2007, vol. I, pp. 193–196.

[19] T. Nakatani, B. H. Juang, T. Yoshioka, K. Kinoshita, and M. Miyoshi, "Importance of energy and spectral features in Gaussian source model for speech dereverberation," in *Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust.*, 2007, pp. 299–302.

[20] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, 2008, pp. 85–88.

[21] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 758–764, 2000.

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

[23] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denoising using multichannel linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1791–1801, Aug. 2007.

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

[25] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[26] 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.
[27] H. Wang and F. Itakura, "Dereverberation of speech signals based on sub-band envelope estimation," *IEICE Trans. Fund.*, vol. E74-A, no. 11, pp. 3576–3583, 1991.
[28] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.
[29] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. I, pp. 557–560.
[30] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation in short time Fourier transform domain with crossband effect compensation," in *Hands-Free Speech Commun, Mic. Arrays*, 2008, pp. 220–223.
[31] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
[32] A. van den Bos, "The multivariate complex normal distribution—A generalization," *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 537–539, Mar. 1995.
[33] Y. Ephraim, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[34] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[35] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.
[36] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-24, no. 3, pp. 243–248, Jun. 1976.
[37] ASJ Continuous Speech Corpus, Acoustical Society of Japan [Online]. Available: http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

Takuya Yoshioka (M'08) received the B.E. and M.Inf. degrees from Kyoto University, Kyoto, Japan, in 2004 and 2006, respectively.

From 2005 to 2006, he was a Trainee at NTT Communication Science Laboratories, NTT Corporation, Kyoto. After that, he joined the NTT Communication Science Laboratories in 2006 as a Research Staff Member. Since then, he has worked on speech dereverberation. His research interests include speech enhancement, speech recognition robust against noisy environments, F0 estimation, and machine learning.

Mr. Yoshioka is a member of 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 at NTT Communication Science Labs, NTT Corporation, Kyoto. Since joining 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. Since 2005, he has visited the Georgia Institute of Technology, Atlanta, as a Visiting Scholar for a year, and investigated a probabilistic formulation of speech dereverberation with Prof. Juang.

Dr. Nakatani received 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 IEICE, and ASJ.

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

Since joining NTT as a Researcher that year, he has been studying signal processing theory and its application to acoustic technologies. Currently, he is the leader of the Signal Processing Group, Media Information Lab, NTT Communication Science Labs. He is also a Visiting Professor of the Graduate School of Information Science and Technology, Hokkaido University.

Dr. Miyoshi received the 1988 IEEE 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 IEICE, ASJ, and AES.