Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 65698, 15 pages
doi:10.1155/2007/65698

Research Article

Dereverberation by Using Time-Variant Nature of Speech Production System

Takuya Yoshioka, Takafumi Hikichi, and Masato Miyoshi

NTT Communication Science Laboratories, NTT Corporation 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

Received 25 August 2006; Revised 7 February 2007; Accepted 21 June 2007

Recommended by Hugo Van hamme

This paper addresses the problem of blind speech dereverberation by inverse filtering of a room acoustic system. Since a speech signal can be modeled as being generated by a speech production system driven by an innovations process, a reverberant signal is the output of a composite system consisting of the speech production and room acoustic systems. Therefore, we need to extract only the part corresponding to the room acoustic system (or its inverse filter) from the composite system (or its inverse filter). The time-variant nature of the speech production system can be exploited for this purpose. In order to realize the time-variance-based inverse filter estimation, we introduce a joint estimation of the inverse filters of both the time-invariant room acoustic and the time-variant speech production systems, and present two estimation algorithms with distinct properties.

1. INTRODUCTION

Room reverberation degrades speech intelligibility or corrupts the characteristics inherent in speech. Hence, dereverberation, which recovers a clean speech signal from its reverberant version, is indispensable for a variety of speech processing applications. In many practical situations, only the reverberant speech signal is accessible. Therefore, the dereverberation must be accomplished with blind processing.

Let an unknown signal transmission channel from a source to possibly multiple microphones in a room be modeled by a linear time invariant system (to provide a unified description independent of the number of microphones, we refer to a set of signal transmission channel(s) from a source to possibly multiple microphones as a signal transmission channel. The channel from the source to each of the microphones is called a subchannel. A set of signal(s) observed by the microphone(s) is refered to as an observed signal. We also refer to an inverse filter set, which is composed of filters applied to the signal observed by each microphone, as an inverse filter). The observed signal (reverberant signal) is then the output of the system driven by the source signal (clean speech signal). On the other hand, the source signal is modeled as being generated by a time variant autoregressive (AR) system corresponding to an articulatory filter driven by an innovations process [1]. In what follows, for the sake of definiteness, the AR system corresponding to the articulatory filter and the system corresponding to the room's signal transmission channel are refered to as the *speech production system* and the *room acoustic system*, respectively. Then, the observed signal is also the output of the composite system of the speech production and room acoustic systems driven by the innovations process. In order to estimate the source signal, the dereverberation may require the inverse filter of the room acoustic system. Therefore, blind speech dereverberation involves the estimation of the inverse filter of the room acoustic system separately from that of the speech production system under the condition that neither the parameters of the speech production system nor those of the room acoustic system are available.

Several approaches to this problem have already been investigated. One major approach is to exploit the diversity between multiple subchannels of the room acoustic system [2–6]. This approach seems to be sensitive to order misdetection or additive noise since it strongly exploits the isomorphic relation between the subspace formed by the source signal and that formed by the observed signal. The so-called prewhitening technique achieved some positive results [7–10]. It relies on the heuristic knowledge that the characteristics of the low order (e.g., 10th order [8]) linear prediction (LP) residue of the observed signal are largely composed of those of the room acoustic system. Based on this knowledge,

this technique regards the residual signal generated by applying LP to the observed signal as the output of the room acoustic system driven by the innovations process. Then, the inverse filter of the room acoustic system can be obtained by using methods designed for i.i.d. series. Although methods incorporating this technique may be less sensitive to additive noise than the subspace approach, the dereverberation performance remains insufficient since the heuristics is just a crude approximation. Also methods that estimate the source signal directly from the observed signal by exploiting features inherent in speech such as harmonicity [11] or sparseness [12] have been proposed. The source estimate is then used as a reference signal when calculating the inverse filter of the room acoustic system. However, the influence of source estimation errors on the inverse filter estimates remains to be revealed, and a detailed investigation should be undertaken.

As an alternative to the above approach, the time variant nature of the speech production system may help us to obtain the inverse filter of the room acoustic system separately from that of the speech production system. Let us consider the inverse filter of a composite system consisting of speech production and room acoustic systems. The overall inverse filter is composed of the inverse filters of the room acoustic and speech production systems. The inverse filter of the room acoustic system is time invariant while that of the speech production system is time variant. Hence, if it is possible to extract only the time invariant subfilter from the overall inverse filter, we can obtain the inverse filter of the room acoustic system. This time-variance-based approach was first proposed by Spencer and Rayner [13] in the context of the restoration of gramophone recordings. They implemented this approach simply; the overall inverse filter is first estimated, and then, it is decomposed into time invariant and time variant subfilters. However, it would be extremely difficult to obtain an accurate estimate of the overall inverse filter, which has both time invariant and time variant zeros especially when the sum of the orders of both systems is large [14]. Therefore, the method proposed in [13] is inapplicable to a room environment.

This paper proposes estimating both the time invariant and time variant subfilters of the overall inverse filter directly from the observed signal. The proposed approach skips the estimation of the overall inverse filter, which is the drawback of the conventional method. Let us consider filtering the observed signal with a time invariant filter and then with a time variant filter. When the output signal is equalized with the innovations process, the time invariant filter becomes the inverse filter of the room acoustic system whereas the time variant filter negates the speech production system. Thus, we can obtain the inverse filter of the room acoustic system simply by adjusting the parameters of the time invariant and time variant filters so that the output signal is equalized with the innovations process. We then propose two blind processing algorithms based on this idea. One uses a criterion involving the second-order statistics (SOS) of the output; the other utilizes the higher-order statistics (HOS). Since SOS estimation demands a relatively small sample size, the SOS-based algorithm will be efficient in terms of the length of the observed signals. On the other hand, the HOS-based algorithm will provide highly accurate inverse filter estimates because the HOS brings additional information. Performance comparisons revealed that the SOS-based algorithm improved the rapid speech transmission index (RASTI), which is a measure of speech intelligibility, from 0.77 to 0.87 by using observed signals of at most five seconds. In contrast, the HOS-based algorithm estimated the inverse filters with a RASTI of nearly one when observed signals of longer than 20 seconds were available. The main variables used in this paper are listed in Table 1 as a reference.

2. PROBLEM STATEMENT

2.1. Problem formulation

The problem of speech dereverberation is formulated as follows. Let a source signal (clean speech signal) be represented by $s(n)$, and the impulse response of an $M\times 1$ linear finite impulse response (FIR) system (room acoustic system) of order K by $\{\mathbf{h}(k) = [h_1(k),\ldots,h_M(k)]^T\}_{0\le k\le K}$. Superscript T indicates the transposition of a vector or a matrix. An observed signal (reverberant signal) $\mathbf{x}(n) = [x_1(n),\ldots,x_M(n)]^T$ can be modeled as

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

Here, $\mathbf{x}(n)$ consists of M signals from the M microphones. By using the transfer function of the room acoustic system, we can rewrite (1) as

$$\mathbf{x}(n) = [\mathbf{H}(z)]s(n), \tag{2}$$

$$\mathbf{H}(z) = \sum_{k=0}^{K} \mathbf{h}(k)z^{-k} = [H_1(z),\ldots,H_M(z)]^T, \tag{3}$$

where $[z^{-1}]$ represents a backward shift operator. $H_m(z)$ is the transfer function of the subchannel of $\mathbf{H}(z)$, corresponding to the signal transmission channel from the source to the mth microphone. Then, the task of dereverberation is to recover the source signal from N samples of the observed signal. This is achieved by filtering the observed signal $\mathbf{x}(n)$ with the inverse filter of the room acoustic system $\mathbf{H}(z)$. Let $y(n)$ denote the recovered signal and let $\{\mathbf{g}(k) = [g_1(k),\ldots,g_M(k)]^T\}_{-\infty\le k\le\infty}$ be the impulse response of the inverse filter. Then, $y(n)$ is represented as

$$y(n) = \sum_{k=\infty}^{\infty} \mathbf{g}(k)^T\mathbf{x}(n-k), \tag{4}$$

or equivalently,

$$y(n) = [\mathbf{G}(z)^T]\mathbf{x}(n), \tag{5}$$

$$\mathbf{G}(z) = \sum_{k=\infty}^{\infty} \mathbf{g}(k)z^{-k}. \tag{6}$$

Note that, by definition, the recovered signal $y(n)$ is a single signal. We want to set up the tap weights $\{g_m(k)\}_{1\le m\le M,\,-\infty\le k\le\infty}$ of the inverse filter so that $y(n)$ is

Table 1: List of main variables.

Variable	Description
M	Number of microphones
N	Number of samples
K	Order of room acoustic system
L	Order of inverse filter of room acoustic system
P	Order of speech production system
W	Size of window function
T	Number of time frames
$s(n)$	Source signal
$\mathbf{x}(n)$	Possibly multichannel observed signal
$y(n)$	Estimate of source signal
$e(n)$	Innovations process
$d(n)$	Estimate of innovations process
$\mathbf{h}(k)$	Impulse response of room acoustic system
$\mathbf{g}(k)$	Impulse response of inverse filter of room acoustic system
$b(k,n)$	Parameter of speech production system
$a(k,n)$	Estimate of parameter of speech production system
$\mathbf{H}(z)$, and so on	Transfer function of room acoustic system $\{\mathbf{h}(k)\}_{0\le k\le K}$, and so on
$\mathrm{GCD}\{P_1(z),\dots,P_n(z)\}$	Greatest common divisor of polynomials $P_1(z),\dots,P_n(z)$
$\mathcal{H}(\boldsymbol{\xi})$	Differential entropy of possibly multivariate random variable $\boldsymbol{\xi}$
$\mathcal{J}(\boldsymbol{\xi})$	Negentropy of possibly multivariate random variable $\boldsymbol{\xi}$
$\mathcal{I}(\xi_1,\dots,\xi_n)$	Mutual information between random variables $\xi_1,\dots,\xi_n$
$\mathcal{K}(\xi_1,\dots,\xi_n)$	Correlatedness between random variables $\xi_1,\dots,\xi_n$
$v(\xi)$	Variance of random variable ξ
$\kappa_i(\xi)$	ith-order cumulant of random variable ξ
$\Sigma(\boldsymbol{\xi})$	Covariance matrix of multivariate random variable $\boldsymbol{\xi}$

equalized with the source signal $s(n)$ up to a constant scale and delay. This requirement can also be stated as

$$\mathbf{G}(z)^T\mathbf{H}(z) = \alpha z^{-\beta}, \tag{7}$$

where α and β are constants representing the scale and delay ambiguity, respectively.

Next, the model of the source signal $s(n)$ is given as follows. A speech signal is widely modeled as being generated by a nonstationary AR process [1]. In other words, the speech signal is the output of a speech production system modeled as a time variant AR system driven by an innovations process. Let $\{b(k,n)\}_{n\in\mathbb{Z},\,1\le k\le P}$, where $\mathbb{Z}$ is the set of integers, denote the time dependent parameters of the speech production system of order P and let $e(n)$ denote the innovations process. Then, $s(n)$ is described as

$$s(n) = \sum_{k=1}^{P} b(k,n)s(n-k) + e(n), \tag{8}$$

or equivalently,

$$s(n) = \left[\frac{1}{1-B(z,n)}\right]e(n), \tag{9}$$

$$B(z,n) = \sum_{k=1}^{P} b(k,n)z^{-k}. \tag{10}$$

In this paper, we assume that

(1) the innovations $\{e(n)\}_{n\in\mathbb{Z}}$ consist of zero-mean independent random variables,
(2) the speech production system $1/(1-B(z,n))$ has no time invariant pole. This assumption is equivalent to the following equation:

$$\mathrm{GCD}\left\{\dots, 1-B(z,0), 1-B(z,1),\dots\right\} = 1, \tag{11}$$

where $\mathrm{GCD}\{P_1(z),\dots,P_n(z)\}$ represents the greatest common divisor of polynomials $P_1(z),\dots,P_n(z)$.

Although assumption (1) does not hold for a voiced portion of speech in a strict sense due to the periodic nature of vocal cord vibration, the assumption has been widely accepted in many speech processing techniques including the linear predictive coding of a speech signal. A comment on the validity of assumption (2) is provided in Section 4.

2.2. Fundamental problem

Figure 1 depicts the system that produces the observed signal from the innovations process. We can see that the observed signal is the output of $\mathbf{H}(z)/(1-B(z,n))$, which we call the *overall acoustic system*, driven by the innovations process.

As mentioned above, our objective is to estimate the inverse filter of $\mathbf{H}(z)$. Despite this objective, we know only the statistical property of the innovations process $e(n)$, specified

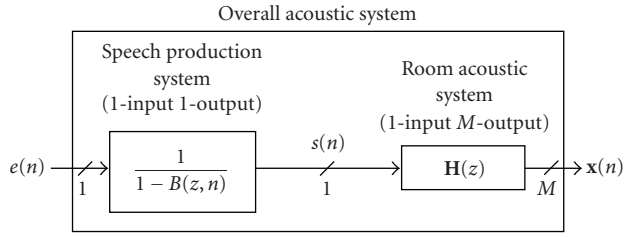

Figure 1: Schematic diagram of system producing observed signal from innovations process.

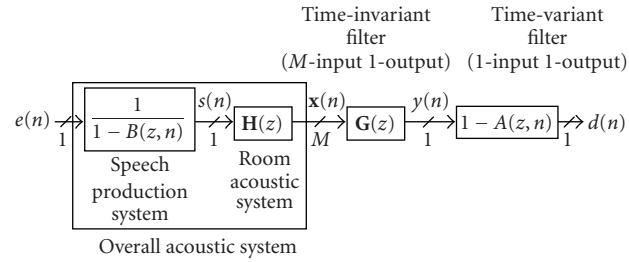

Figure 2: Schematic diagram of global system from innovations process to its estimate.

by assumption (1); neither the parameters of $1/(1 - B(z,n))$ nor those of $\mathbf{H}(z)$ are available. Therefore, we face the critical problem of how to obtain the inverse filter of $\mathbf{H}(z)$ separately from that of $1/(1 - B(z,n))$ with blind processing. This is the cause of the so-called excessive whitening problem [6], which indicates that applying methods designed for i.i.d. series (e.g., see [15, 16] and references therein) to a speech signal results in cancelling not only the characteristics of the room acoustic system $\mathbf{H}(z)$ but also the average characteristics of the speech production system $1/(1 - B(z,n))$.

3. TIME-VARIANCE-BASED APPROACH

In order to overcome the problem mentioned above, we have to exploit a characteristic that differs for the room acoustic system $\mathbf{H}(z)$ and the speech production system $1/(1 - B(z,n))$. We use the time variant nature of the speech production system as such a characteristic.

Let us consider the inverse filter of the overall acoustic system $\mathbf{H}(z)/(1 - B(z,n))$. Since the overall acoustic system consists of a time variant part $1/(1 - B(z,n))$ and a time invariant part $\mathbf{H}(z)$, the inverse filter accordingly has both time invariant and time variant zeros. The set of time invariant zeros forms the inverse filter of the room acoustic system $\mathbf{H}(z)$ while the time variant zeros constitute the inverse filter of the speech production system $1/(1 - B(z,n))$. Hence, we can obtain the inverse filter of the room acoustic system by extracting the time invariant subfilter from the inverse filter of the overall acoustic system.

3.1. Review of conventional methods

A method of implementing the time-variance-based inverse filter estimation is proposed in [13, 17]. The method proposed in [13, 17] identifies the speech production system and the room acoustic system assuming that both systems are modeled as AR systems. The overall acoustic system is first estimated from several contiguous disjoint observation frames. In this step, it is assumed that the overall acoustic system is time invariant within each frame. Then, poles commonly included in the framewise estimates of the overall acoustic system are collected to extract the time invariant part of the overall acoustic system.

The method imposes the following two conditions.

(i) The frame size is larger than the order of the room acoustic system as well as that of the speech production system.

(ii) None of the system parameters change within a single frame.

However, the parameters of the speech production system change by tens of milliseconds while the order of the room acoustic system may be equivalent to several hundred milliseconds. Therefore, we can never design a frame size that meets those two conditions. This frame-size problem is discussed in more detail in Section 3.2.

Moreover, this method assumes that the room acoustic system is minimum phase, which may be an unrealistic assumption. Therefore, it is difficult to apply this method to an actual room environment.

Reference [14] proposes another method of implementing the time-variance-based inverse filter estimation. The method estimates only the room acoustic system based on maximum a posteriori estimation assuming that the innovations process $e(n)$ is Gaussian white noise. However, the method also assumes the room acoustic system to be minimum phase.

3.2. Novel method based on joint estimation of time invariant/time variant subfilters

The two requirements for the frame size with the conventional method arise from the fact that it estimates the overall acoustic system in the first step. Therefore, we propose the joint estimation of the time invariant and time variant subfilters of the inverse filter of the overall acoustic system directly from the observed signal $\mathbf{x}(n)$.

Let us consider filtering $\mathbf{x}(n)$ with time invariant filter $\mathbf{G}(z)$ and then with time variant filter $1 - A(z,n)$ (see Figure 2). If we represent the parameters of $1 - A(z,n)$ by $\{a(k,n)\}_{1 \le k \le P}$, the final output $d(n)$ is given as follows:

$$d(n) = y(n) - \sum_{k=1}^{P} a(k,n)\, y(n-k), \tag{12}$$

or equivalently,

$$d(n) = [1 - A(z,n)]y(n), \tag{13}$$

$$A(z,n) = \sum_{k=1}^{P} a(k,n)z^{-k}, \tag{14}$$

where $y(n)$ is given by (5). Then, we have the following theorem under assumption (2).

Theorem 1. *Assume that the final output signal $d(n)$ is equalized with innovations process $e(n)$ up to a constant scale and delay, and that $1 - A(z,n)$ has no time invariant zero:*

$$d(n) = \alpha e(n - \beta), \tag{15}$$

$$\mathrm{GCD}\{1 - A(z,1), \ldots, 1 - A(z,N)\} = 1. \tag{16}$$

Then, the time invariant filter $\mathbf{G}(z)$ satisfies (7).

Proof. The proof is given in Appendix A. □

This theorem states that we simply have to set up the tap weights $\{g_m(k)\}$[1] and $\{a(k,n)\}$ so that $d(n)$ is equalized with $\alpha e(n - \beta)$. The calculated time invariant filter $\mathbf{G}(z)$ corresponds to the inverse filter of the room acoustic system $\mathbf{H}(z)$, and the time variant filter $1 - A(z,n)$ corresponds to that of the speech production system $1/(1 - B(z,n))$. Thus, we can conclude that the joint estimation of the time invariant/time variant subfilters is a possible solution to the problem described in Section 2.2.

At this point, we can clearly explain the drawback of the conventional method with a large frame size. When using a large frame size, it is impossible to completely equalize $d(n)$ with $\alpha e(n - \beta)$ because $1/(1 - B(z,n))$ varies within a single frame. Hence, the estimate of the overall acoustic system in each frame is inevitably contaminated by estimation errors. These errors make it difficult to extract static poles from the framewise estimates of the overall acoustic system. By contrast, the joint estimation that we propose does not involve the estimation of the inverse filter of the overall acoustic system. Therefore, a frame size shorter than the order of the room acoustic system can be employed, which enables us to equalize $d(n)$ with $\alpha e(n - \beta)$.

Since the innovations process $e(n)$ is inaccessible in reality, we have to develop criteria defined solely by using $d(n)$. These criteria are provided in the next two sections. The algorithms derived can deal with a nonminimum phase system as the room acoustic system since they use multiple microphones and/or the HOS of the output $d(n)$ [15, 16].

4. ALGORITHM USING SECOND-ORDER STATISTICS

Since output signal $d(n)$ is an estimate of innovations process $e(n)$, it would be natural to set up the tap weights $\{g_m(k)\}$ and $\{a(k,n)\}$ so that the statistical property of the outputs $\{d(n)\}_{1 \le n \le N}$ satisfies assumption (1). In this section, we develop a criterion based only on the SOS of $\{d(n)\}$. To be more precise, we try to uncorrelate $\{d(n)\}$.

We assume the following two conditions additionally in this section.

(i) $M \ge 2$, that is, we use multiple microphones.
(ii) Subchannel transfer functions $H_1(z), \ldots, H_M(z)$ have no common zero.

Under these assumptions, the observed signal $\mathbf{x}(n)$ is an AR process driven by the source signal $s(n)$ [16]. Therefore, we can substitute an FIR inverse filter of order L for the doubly-infinite inverse filter in (4) as

$$y(n) = \sum_{k=0}^{L} \mathbf{g}(k)^T \mathbf{x}(n - k). \tag{17}$$

Here, we can restrict the first tap of $\mathbf{G}(z)$ as

$$g_m(0) = \begin{cases} 1 & m = 1, \\ 0 & m = 2, \ldots, M, \end{cases} \tag{18}$$

where the microphone with $m = 1$ is nearest to the source (see [16] for details).

4.1. Loss function

Let $\mathcal{K}(\xi_1, \ldots, \xi_n)$ denote a suitable measure of correlatedness between random variables $\xi_1, \ldots, \xi_n$. Then, the problem is mathematically formulated as

$$\begin{aligned} &\underset{\{a(k,n)\},\{g_m(k)\}}{\text{minimize}} \ \mathcal{K}(d(1), \ldots, d(N)) \\ &\text{subject to } \{1 - A(z,n)\}_{1 \le n \le N} \text{ being minimum phase.} \end{aligned} \tag{19}$$

The constraint of (19) is intended to stabilize the estimate, $1/(1 - A(z,n))$, of the speech production system.

First, we need to define the correlatedness measure $\mathcal{K}(\cdot)$. Several criteria for measuring the correlatedness between random variables have been developed [18, 19]. We use the criterion proposed in [19] since it can be further simplified as described later. The criterion is defined as

$$\mathcal{K}(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \log v(\xi_i) - \log \left| \det \Sigma(\boldsymbol{\xi}) \right|, \tag{20}$$

$$\boldsymbol{\xi} = [\xi_n, \ldots, \xi_1]^T, \tag{21}$$

where $v(\xi_1), \ldots, v(\xi_n)$, respectively, represent the variances of random variables $\xi_1, \ldots, \xi_n$, and $\Sigma(\boldsymbol{\xi})$ denotes the covariance matrix of $\boldsymbol{\xi}$. Definition (20) is a suitable measure of correlatedness in that it satisfies

$$\mathcal{K}(\xi_1, \ldots, \xi_n) \ge 0 \tag{22}$$

with equality if and only if random variables $\xi_1, \ldots, \xi_n$ are uncorrelated as

$$i \ne j \iff E\{\xi_i \xi_j\} = 0, \tag{23}$$

[1] Hereafter, we will omit the range of indices unless necessary.

where $E\{\cdot\}$ denotes an expectation operator. Then, we will try to minimize

$$\mathcal{K}(d(1),\ldots,d(N)) = \sum_{n=1}^{N} \log v(d(n)) - \log\left|\det \Sigma(\mathbf{d})\right|, \tag{24}$$

$$\mathbf{d} = [d(N),\ldots,d(1)]^T \tag{25}$$

with respect to $\{a(k,n)\}$ and $\{g_m(k)\}$. This loss function can be further simplified as follows under (18) (see Appendix B):

$$\mathcal{K}(d(1),\ldots,d(N)) = \sum_{n=1}^{N} \log v(d(n)) + \text{constant}. \tag{26}$$

Hence, problem (19) is finally reduced to

$$\begin{aligned} &\underset{\{a(k,n)\},\,\{g_m(k)\}}{\text{minimize}} \sum_{n=1}^{N} \log v(d(n)) \\ &\text{subject to } \{1 - A(z,n)\} \text{ being minimum phase.} \end{aligned} \tag{27}$$

Therefore, we have to set up tap weights $\{a(k,n)\}$ and $\{g_m(k)\}$ under (18) so as to minimize the logarithmic mean of the variances of outputs $\{d(n)\}$.

Next, we show that the set of $1 - A(z,n)$ and $\mathbf{G}(z)$ that minimizes the loss function of (27) equalizes the output signal $d(n)$ with the innovations process $e(n)$.

Theorem 2. *Suppose that there is an inverse filter, $\mathbf{G}(z)$, of the room acoustic system that satisfies (7) and (18). Then, $\sum_{n=1}^{N} \log v(d(n))$ achieves a minimum if and only if*

$$d(n) = \alpha e(n-\beta) = h_1(0)e(n). \tag{28}$$

Proof. The proof is presented in Appendix C. □

With Theorems 1 and 2, a solution to problem (27) provides the inverse filters of the room acoustic system and the speech production system.

Remark 1. Let us assume that the variance of $d(n)$ is stationary. The loss function of (27) is then equal to $N \log v(d(n))$. Because the logarithmic function is increasing monotonically, the loss function is further simplified to $Nv(d(n))$, which may be estimated by $\sum_{n=1}^{N} d(n)^2$. Thus, the loss function of (27) is equivalent to the traditional least squares (LS) criterion when the variance of $d(n)$ is stationary. However, since the variance of the innovations process indeed changes with time, the loss function of (27) may be more appropriate than the LS criterion. This conjecture will be justified by the experiments described later.

4.2. Algorithm

In this section, we derive an algorithm for accomplishing (27). Before we proceed, we introduce an approximation of time variant filter $1 - A(z,n)$. Since a speech signal within a short time frame of several tens of milliseconds is almost stationary, we approximate $1 - A(z,n)$ by using a filter that is globally time variant but locally time invariant as

$$1 - A(z,n) = 1 - A_i(z), \quad i = \left\lfloor \frac{n-1}{W} + 1 \right\rfloor, \tag{29}$$

where W is the frame size and $\lfloor\cdot\rfloor$ represents the floor function. Under this approximation, $d(n)$ is produced from $y(n)$ as follows. The outputs $\{y(n)\}_{1\le n\le N}$, of $\mathbf{G}(z)$ are segmented into T short time frames by using a W-sample rectangular window function. This generates T segments $\{y(n)\}_{N_1\le n\le N_1+W-1},\ldots,\{y(n)\}_{N_T\le n\le N_T+W-1}$, where N_i is the first index of the ith frame satisfying $N_1 = 1, N_T+W-1 = N$, and $N_i + W = N_{i+1}$. Then, $y(n)$ in the ith frame is processed through $1 - A_i(z)$ to yield $d(n)$ as

$$d(n) = y(n) - \sum_{k=1}^{P} a_i(k)\, y(n-k). \tag{30}$$

By using this approximation, problem (27) is reformulated as

$$\begin{aligned} &\underset{\{a_i(k)\}_{1\le i\le T,\,1\le k\le P},\,\{g_m(k)\}_{1\le m\le M,\,1\le k\le L}}{\text{minimize}} \sum_{n=1}^{N} \log v(d(n)) \\ &\text{subject to } \{1 - A_i(z)\}_{1\le i\le T} \text{ being minimum phase.} \end{aligned} \tag{31}$$

We solve problem (31) by employing an alternating variables method. The method minimizes the loss function with respect first to $\{a_i(k)\}$ for fixed $\{g_m(k)\}$, then to $\{g_m(k)\}$ for fixed $\{a_i(k)\}$, and so on. Let us represent the fixed value of $g_m(k)$ by $\tilde{g}_m(k)$ and that of $a_i(k)$ by $\tilde{a}_i(k)$. Then, we can formulate the optimization problems for estimating $\{a_i(k)\}$ and $\{g_m(k)\}$ as

$$\begin{aligned} &\underset{\{a_i(k)\}_{1\le i\le T,\,1\le k\le P}}{\text{minimize}} \sum_{n=1}^{N} \log v(d(n))\Big|_{\{g_m(k)\}=\{\tilde{g}_m(k)\}} \\ &\text{subject to } \{1 - A_i(z)\} \text{ being minimum phase,} \end{aligned} \tag{32}$$

$$\underset{\{g_m(k)\}_{1\le m\le M,\,1\le k\le L}}{\text{minimize}} \sum_{n=1}^{N} \log v(d(n))\Big|_{\{a_i(k)\}=\{\tilde{a}_i(k)\}}. \tag{33}$$

Note that only $\{g_m(k)\}$ with $k \ge 1$ are adjusted. The first tap weights $\{g_m(0)\}$ are fixed as (18). By repeating the optimization cycle of (32) and (33) R_1 times, we obtain the final estimates of $a_i(k)$ and $g_m(k)$.

First, let us derive the algorithm that accomplishes (32). We first note that (32) is achieved by solving the following problem for each frame number i:

$$\begin{aligned} &\underset{\{a_i(k)\}_{1\le k\le P}}{\text{minimize}} \sum_{n=N_i}^{N_i+W-1} \log v(d(n))\Big|_{\{g_m(k)\}=\{\tilde{g}_m(k)\}} \\ &\text{subject to } 1 - A_i(z) \text{ being minimum phase.} \end{aligned} \tag{34}$$

Let us assume that $d(n)$ is stationary within a single frame. Then, the loss function of (34) becomes

$$\sum_{n=N_i}^{N_i+W-1} \log v(d(n)) = N \log v(d(n)). \tag{35}$$

Furthermore, because of the monotonically increasing property of the logarithmic function, the loss function becomes equivalent to $Nv(d(n))$, which can be estimated by $\sum_{n=N_i}^{N_i+W-1} d(n)^2$. Thus, the solution to (34) is obtained by minimizing the mean square of $d(n)$. Such a solution is calculated by applying linear prediction (LP) to $\{y(n)\}_{N_i \le n \le N_i+W-1}$. It should be noted that LP guarantees that $1 - A_i(z)$ is minimum phase when the autocorrelation method is used [1].

Next, we derive the algorithm to solve (33). We realize (33) by using the gradient method. By calculating the derivative of loss function $\sum_{n=1}^{N} \log v(d(n))$, we obtain the following algorithm (see Appendix D for the derivation):

$$g_m(k)' = g_m(k) + \delta \sum_{i=1}^{T} \frac{\langle d(n) v_{m,i}(n-k) \rangle_{n=N_i}^{N_i+W-1}}{\langle d(n)^2 \rangle_{n=N_i}^{N_i+W-1}}, \tag{36}$$

$$v_{m,i}(n) = x_m(n) - \sum_{k=1}^{P} a_i(k) x_m(n-k), \tag{37}$$

where $\langle \cdot \rangle_{n=N_i}^{N_i+W-1}$ is an operator that takes an average from N_ith to $(N_i + W - 1)$th samples, and δ is the step size. The update procedure (36) is repeated R_2 times. Since the gradient-based optimization of $\{g_m(k)\}$ is involved in each (32)-(33) optimization cycle, (36) is performed $R_1 R_2$ times in total.

Remark 2. Now, let us consider the special case of $R_1 = 1$. Assume that we initialize $\{g_m(k)\}$ as

$$g_m(k) = 0, \quad 1 \le \forall m \le M,\ 1 \le \forall k \le L. \tag{38}$$

Then, $\{a_i(k)\}$ is estimated via LP directly from the observed signal, and $\{g_m(k)\}$ is estimated by using those estimates of $\{a_i(k)\}$. This is essentially equivalent to methods that use the prewhitening technique [7–10]. In this way, the prewhitening technique, which has been used heuristically, is derived from the models of source and room acoustics explained in Section 2. Moreover, by repeating the (32)-(33) cycle, we may obtain more precise estimates.

4.3. Experimental results

We conducted experiments to demonstrate the performance of the algorithm described above. We took Japanese sentences uttered by 10 speakers from the ASJ-JNAS database [20]. For each speaker, we made signals of various lengths by concatenating his or her utterances. These signals were used as the source signals, and by using these signals, we could investigate the dependence of the performance on the signal length. The observed signals were simulated by convolving the source signals with impulse responses measured in a room. The room layout is illustrated in Figure 3. The order of the impulse responses, K, was 8000. The reverberation time was around 0.5 seconds. The signals were all sampled at 8 kHz and quantized with 16-bit resolution.

The parameter settings are listed in Table 2. The initial estimates of the tap weights were set as

$$g_m(k) = 0, \quad 1 \le \forall m \le M,\ 1 \le \forall k \le L \tag{39}$$

while $\{g_m(0)\}_{1 \le m \le M}$ are fixed as (18).

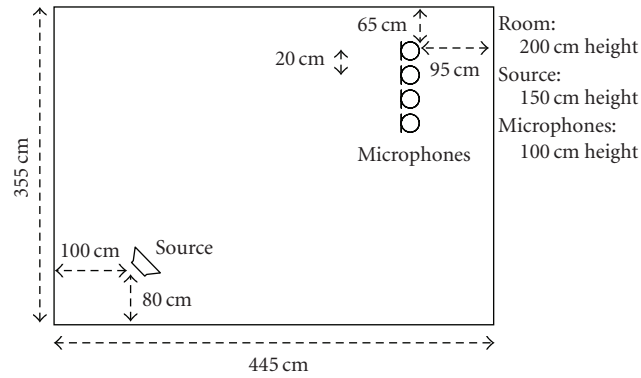

Figure 3: Room layout.

Table 2: Parameter settings. Each optimization (32) is realized by LP whereas each (33) is implemented by repeating (36).

Number of microphones	M	4
Order of $\mathbf{G}(z)$	L	1000
Frame size	W	200
Order of $A_i(z)$	P	16
Number of repetitions of (32)-(33) cycle	R_1	6
Number of repetitions of (36)	R_2	50

Offline experiments were conducted to evaluate the fundamental performance. For each speaker and signal length, the inverse filter was estimated by using the corresponding observed signal. The estimated inverse filter was applied to the observed signal to calculate the accuracy of the estimate. Finally, for each signal length, we averaged the accuracies over all the speakers to obtain plots such as those in Figure 4. In Figure 4, the horizontal axis represents the signal length, and the vertical axis represents the averaged accuracy, whose measures are explained below.

Since the proposed algorithm estimates the inverse filters of the room acoustic system and the speech production system, we accordingly evaluated the dereverberation performance by using two measures. One was the rapid speech transmission index (RASTI[2]) [21], which is the most common measure for quantifying speech intelligibility from the viewpoint of room acoustics. We used RASTI as a measure for evaluating the accuracy of the estimated inverse filter of the room acoustic system. According to [21], RASTI is defined based on the modulation transfer function (MTF), which quantifies the flattening of power fluctuations by reverberation. A RASTI score closer to one indicates higher speech intelligibility. The other is the spectral distortion (SD) [22] between the speech production system $1/(1 - B(z, n))$ and its estimate $1/(1 - A(z, n + \beta))$. Since the characteristics of the speech production system can be regarded as those of

[2] We used RASTI instead of the speech transmission index (STI) [21], which is the precise version of RASTI, because calculating an STI score requires a sampling frequency of 16 kHz or greater.

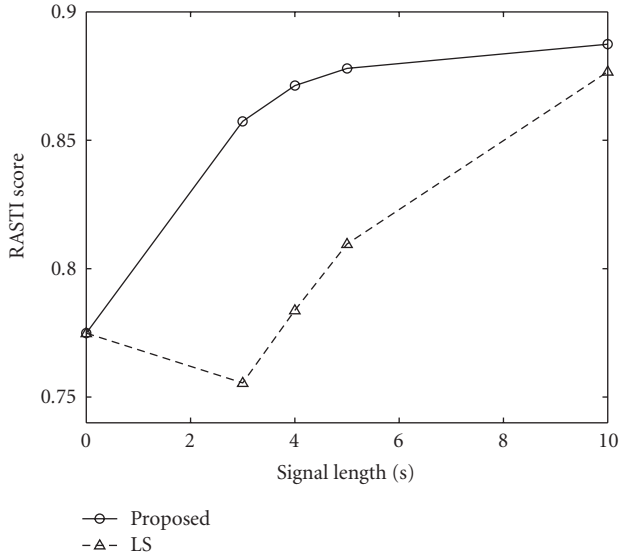

Figure 4: RASTI as a function of observed signal length.

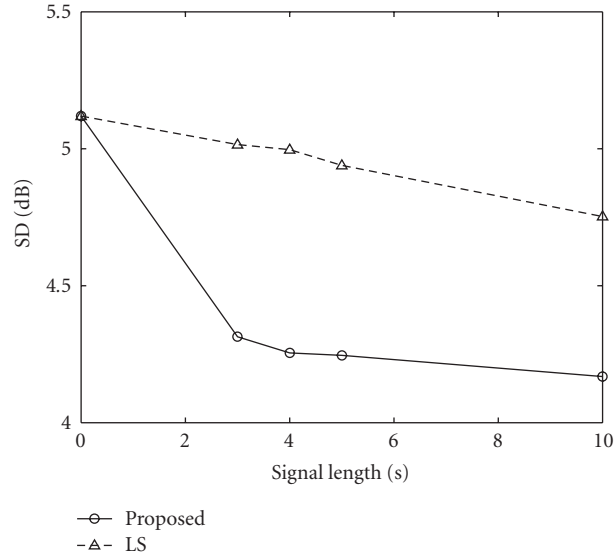

Figure 5: SD as a function of observed signal length.

the clean speech signal, the SD represents the extraction error of the speech characteristics. We used the SD as a measure for assessing the accuracy of the estimated inverse filter of the speech production sytem. The reference $1/(1 - B(z, n))$ was calculated by applying LP to the clean speech signal $s(n)$ segmented in the same way as the recovered signal $y(n)$.

To show the effectiveness of incorporating the nonstationarity of the innovations process (see the remark in the last paragraph of Section 4.1), we compared the performance of the proposed algorithm with that of an algorithm based on the least squares (LS) criterion. The LS-based algorithm solves

$$\underset{\{a_i(k)\},\{g_m(k)\}}{\text{minimize}} \sum_{n=1}^{N} d(n)^2 \quad \text{subject to } \{1 - A_i(z)\} \text{ being minimum phase.} \tag{40}$$

Such an algorithm can be easily obtained by replacing the algorithm solving (33) by the multichannel LP [16, 23].

Figure 4 shows the RASTI score averaged over the 10 speakers' results as a function of the length of the observed signal. Figure 5 shows the SD averaged over the results for all time frames and speakers. There was little difference between the results of the proposed algorithm and those of the LS-based algorithm when the length of the observed signal was above 10 seconds. Hence, we plot the results for observed signals duration up to 10 seconds in Figures 4 and 5 to highlight the difference between the two algorithms. We can see that the proposed algorithm outperformed the algorithm based on the LS criterion especially when the observed signals were short.

We found that, among the 10 speakers, the dereverberation performance for the male speakers was a bit better than that for the female speakers. This is probably because assumption (1) fits better for male speakers because the pitches of male speeches are generally lower than those of female speeches.

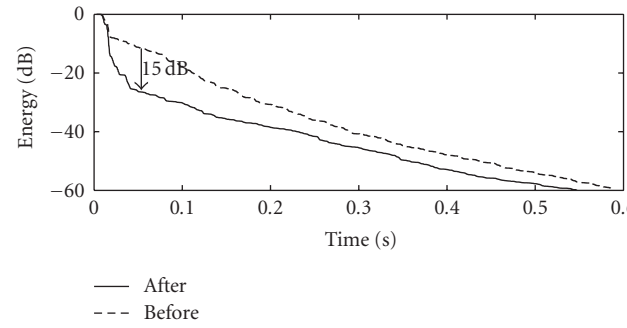

Figure 6: Energy decay curves of impulse responses before and after dereverberation.

In Figure 6, we show examples of the energy decay curves of impulse responses before and after the dereverberation obtained by using an observed signal of five seconds. A clear reduction in reflection energy can be seen; there was a 15 dB reduction in the reverberant energy 50 milliseconds after the arrival of the direct sound.

From the above results, we conclude that the proposed algorithm can estimate the inverse filter of the room acoustic system with a relatively short 3–5 second observed signal.

5. ALGORITHM USING HIGHER-ORDER STATISTICS

In this section, we derive an algorithm that estimates $\{a(k, n)\}_{1\le n\le N, 1\le k\le P}$ and $\{g_m(k)\}_{1\le m\le M, 0\le k\le L}$ so that the outputs $\{d(n)\}_{1\le n\le N}$ become statistically independent of each other. Statistical independence is a stronger requirement than the uncorrelatedness exploited by the algorithm described in the preceding section since the independence of

random variables is characterized by both their SOS and their HOS. Therefore, an algorithm based on the independence of $\{d(n)\}$ is expected to realize a highly accurate inverse filter estimation because it fully uses the characteristics of the innovations process specified by assumption (1).

Before presenting the algorithm, we formulate a theorem about the uniqueness of the estimates, $\{d(n)\}$, of the innovations $\{e(n)\}$. In this section, we also assume that

(i) the innovations $\{e(n)\}$ have non-Gaussian distributions,
(ii) the innovations $\{e(n)\}$ satisfy the Lindeberg condition [24].

Under these assumptions, we have the following theorem.

Theorem 3. *Suppose that variables $\{d(n)\}$ are not deterministic. If $\{d(n)\}$ are statistically independent with non-Gaussian distributions, then $d(n)$ is equalized with $e(n)$ except for a possible scaling and delay.*

Proof. The proof is deferred to Appendix E. □

By using Theorems 1 and 3, it is clear that the inverse filters of the room acoustic system and the speech production system are uniquely identifiable.

In practice, the doubly-infinite inverse filter $\mathbf{G}(z)$ in (4) is approximated by the L-tap FIR filter as

$$y(n) = \sum_{k=0}^{L} \mathbf{g}(k)^T \mathbf{x}(t-k). \tag{41}$$

Unlike the SOS-based algorithm, we need not constrain the first tap weights as (18). Thus, we estimate $\{g_m(k)\}$ with $k \geq 0$ in this section.

5.1. Loss function

Let us represent the mutual information of random variables $\xi_1, \ldots, \xi_n$ by $\mathcal{I}(\xi_1, \ldots, \xi_n)$. By using the mutual information as a measure of the interdependence of the random variables, we minimize the loss function defined as $\mathcal{I}(d(1), \ldots, d(N))$ with respect to $\{a(k,n)\}$ and $\{g_m(k)\}$ under the constraint that instantaneous systems $\{1 - A(z,n)\}$ are minimum phase in a similar way to (19). The loss function can be rewritten as (see Appendix F)

$$\mathcal{I}(d(1), \ldots, d(N)) = -\sum_{n=1}^{N} \mathcal{J}(d(n)) + \mathcal{K}(d(1), \ldots, d(N)), \tag{42}$$

where $\mathcal{J}(\xi)$ denotes the negentropy [25] of random variable ξ. The computational formula of the negentropy is given later. The negentropy represents the nongaussianity of a random variable. From (42), what we try to solve is formulated as

$$\begin{aligned} &\underset{\{a(k,n)\},\{g_m(k)\}}{\text{minimize}} \left(-\sum_{n=1}^{N} \mathcal{J}(d(n)) + \mathcal{K}(d(1), \ldots, d(N)) \right) \\ &\text{subject to } \{1 - A(z,n)\} \text{ being minimum phase.} \end{aligned} \tag{43}$$

By comparing (43) with (19), it is found that (43) exploits the negentropies of $\{d(n)\}$ in addition to the correlatedness between $\{d(n)\}$ as a criterion. Therefore, we try not only to uncorrelate outputs $\{d(n)\}$ but also to make the distributions of $\{d(n)\}$ as far from the Gaussian as possible.

5.2. Algorithm

As regards time variant filter $1 - A(z,n)$, we again use approximation (29). Then, we solve

$$\begin{aligned} &\underset{\{a_i(k)\},\{g_m(k)\}}{\text{minimize}} \left(-\sum_{n=1}^{N} \mathcal{J}(d(n)) + \mathcal{K}(d(1), \ldots, d(N)) \right) \\ &\text{subject to } \{1 - A_i(z)\} \text{ being minimum phase} \end{aligned} \tag{44}$$

instead of (43).

Problem (44) is solved by the alternating variables method in a similar way to the algorithm in Section 4. Namely, we repeat the minimization of the loss function with respect to $\{a_i(k)\}$ for fixed $\{g_m(k)\}$ and minimization with respect to $\{g_m(k)\}$ for fixed $\{a_i(k)\}$. However, since the loss function of (44) is very complicated, we derive a suboptimal algorithm by introducing the following assumptions found in our preliminary experiment.

(i) Given $\{g_m(k)\}$, or equivalently, given $y(n)$, the set of parameters $\{a_i(k)\}$ that minimizes $\mathcal{K}(d(1), \ldots, d(N))$ also reduces the loss function of (44).
(ii) Given $\{a_i(k)\}$, the set of parameters $\{g_m(k)\}$ that minimizes $(-\sum_{n=1}^{N} \mathcal{J}(d(n)))$ also reduces the loss function of (44).

With assumption (i), we again estimate $\{a_i(k)\}_{1 \leq k \leq P}$ by applying LP to segment $\{y(n)\}_{N_i \leq n \leq N_i + W - 1}$, which is the output of $\mathbf{G}(z)$, for each i. It should be remembered that we can obtain minimum-phase estimates of $\{1 - A_i(z)\}$ by using LP.

Next, we estimate $\{g_m(k)\}$ for fixed $\{a_i(k)\}$ by maximizing $\sum_{n=1}^{N} \mathcal{J}(d(n))$ based on assumption (ii). By using the Gram-Charlier expansion and retaining dominant terms, we can approximate the negentropy $\mathcal{J}(\xi)$ of random variable ξ as [26]

$$\mathcal{J}(\xi) \simeq \frac{\kappa_3(\xi)^2}{12 v(\xi)^3} + \frac{\kappa_4(\xi)^2}{48 v(\xi)^4}, \tag{45}$$

where $\kappa_i(\xi)$ represents the ith order cumulant of ξ. Generally, the innovations of a speech signal have supergaussian distributions whose third-order cumulants are negligible compared with its fourth-order cumulants. Therefore, we finally reach the following problem in the estimation of $\{g_m(k)\}$:

$$\begin{aligned} &\underset{\{g_m(k)\}_{1 \leq m \leq M, 0 \leq k \leq L}}{\text{maximize}} \sum_{n=1}^{N} \frac{\kappa_4(d(n))}{v(d(n))^2} \Bigg|_{\{a_i(k)\} = \{\widetilde{a}_i(k)\}} \\ &\text{subject to } \sum_{m=1}^{M} \sum_{k=0}^{L} g_m(k)^2 = 1. \end{aligned} \tag{46}$$

We again note that the range in k is from 0 to L unlike (33). The constraint of (46) is intended to determine the constant

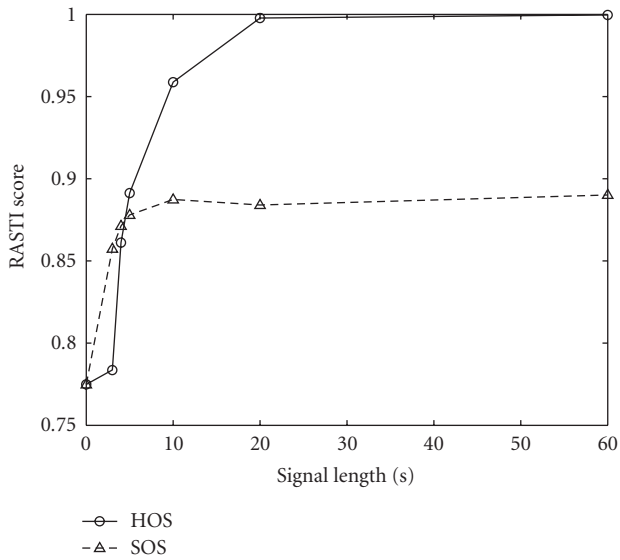

Figure 7: RASTI as a function of observed signal length.

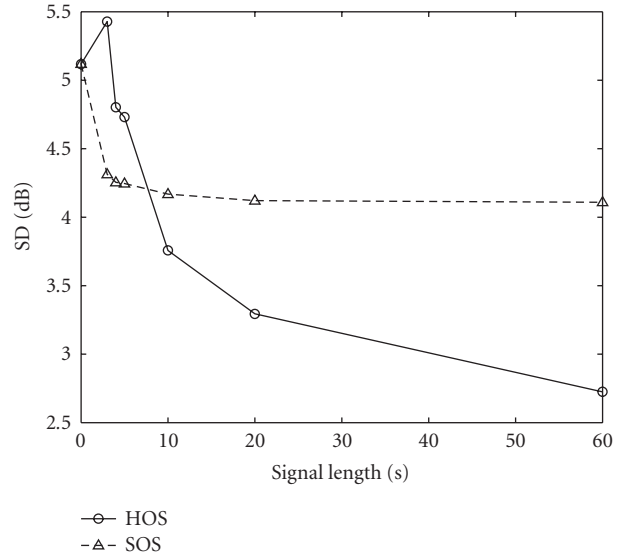

Figure 8: SD as a function of observed signal length.

scale α arbitrarily. We use the gradient method to realize this maximization. By taking the derivative of the loss function of (46), we have the following algorithm:

$$\begin{aligned} g_m(k)' &= g_m(k) \\ &\quad + \delta \sum_{i=1}^{T} \frac{4}{\langle d(n)^2 \rangle^4} \\ &\quad \times \big(\langle d(n)^3 v_{m,i}(n-k) \rangle \langle d(n)^2 \rangle^2 \\ &\qquad - \langle d(n)^4 \rangle \langle d(n)^2 \rangle \langle d(n) v_{m,i}(n-k) \rangle \big), \end{aligned} \tag{47}$$

$$g_m(k)'' = \frac{g_m(k)'}{\sum_{m=1}^{M} \sum_{k=0}^{L} g_m(k)'^2},$$

where the averages are calculated for indices N_i to $N_i + W - 1$. Here, we have again used the assumption that $d(n)$ is stationary within a single frame just as we did in the derivation of (36).

Remark 3. While we can easily estimate $\{a_i(k)\}$ and $\{g_m(k)\}$ with assumptions (i) and (ii), the convergence of the algorithm is not guaranteed because the assumptions may not always be true. We examine this issue experimentally. It is hoped that future work will reveal the theoretical background to the assumptions.

5.3. Experimental results

We compared the dereverberation performance of the HOS-based algorithm proposed in this section with that of the SOS-based algorithm described in the previous section. We used the same experimental setup as that in the previous section except for the iteration parameters R_1 and R_2, which we set at 10 and 20, respectively.

Figure 7 shows the RASTI score averaged over the 10 speakers' results as a function of the length of the observed signal. As expected, we can see that the HOS-based algorithm outperformed the SOS-based algorithm when the observed signal was relatively long. In particular, when an observed signal of longer than 20 seconds was available, the RASTI score was nearly equal to one. Figure 8 shows the average SD. Again, we can confirm the great superiority of the HOS-based algorithm to the SOS-based algorithm in terms of asymptotic performance.

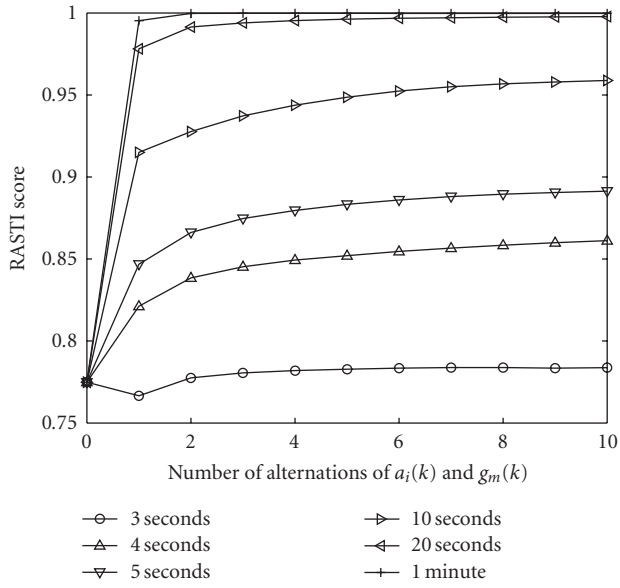

Figure 9: RASTI as a function of iteration number.

In Figure 9, we plot the average RASTI score as a function of the number of alternations of estimation parameters $\{a_i(k)\}$ and $\{g_m(k)\}$. We can clearly see the convergence

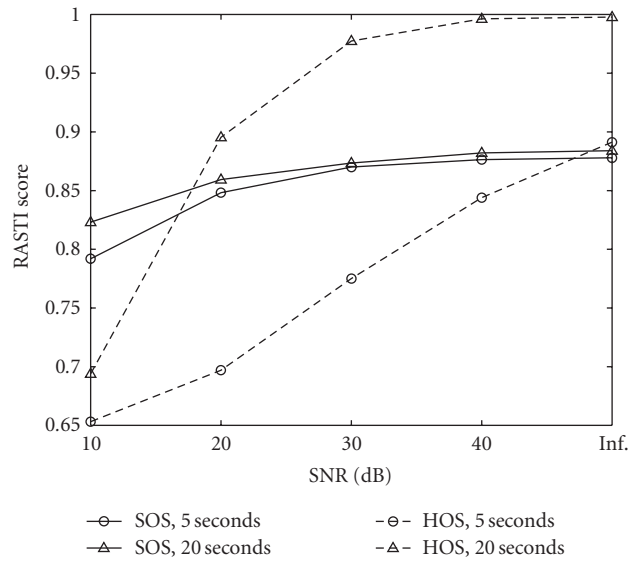

Figure 10: RASTI obtained in the presence of noise.

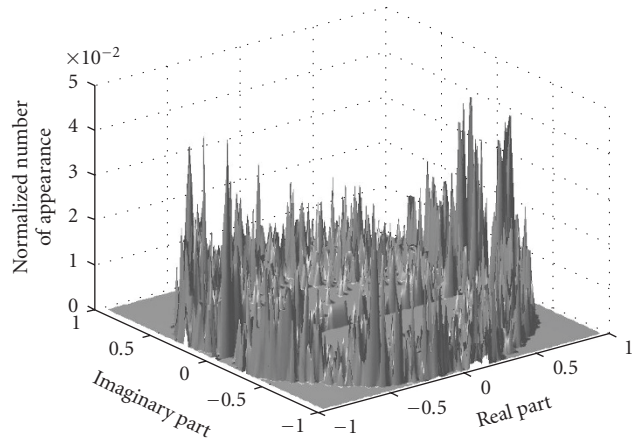

Figure 11: Histogram showing the number of poles of the speech production system in each small region in the complex plane.

of the RASTI score. The RASTI score converges particularly rapidly when the observed signal length is sufficiently large.

6. DISCUSSION

6.1. Effect of additive noise

Thus far, we have considered a system without any additive noise. In this section, we experimentally examine the effect of additive noise on the performance of the proposed algorithms[3].

We tested a case where the observed signal was contaminated by additive white Gaussian noise with signal to noise ratios (SNR) of 40, 30, 20, and 10 dB. Since the proposed methods do not involve noise reduction, we measured the performance as a RASTI score calculated by using the impulse response of equalized room acoustic system $\mathbf{G}(z)^T\mathbf{H}(z)$.

In Figure 10, we plot the average RASTI scores as a function of the SNR for observed signals of five and twenty seconds. The SOS-based algorithm was relatively robust against additive noise. Although the performance of the HOS-based algorithm was degraded more severely than that of the SOS-based algorithm, the former still exhibited excellent performance in the presence of noise with an SNR of 30 dB or greater when the observed signal was 20 seconds long.

Thus, it is a promising way to combine the proposed algorithms with traditional noise reduction methods such as spectral subtraction [28] in a noisy environment with a severe SNR. An investigation of such a combination is however beyond the scope of this paper.

6.2. Validity of assumption (2)

Assumption (2) is one of the essential assumptions that form the basis of the proposed algorithms. Here we investigate its validity.

Figure 11 is an example histogram showing the number of poles of the speech production system included in a clean speech signal of five seconds in each small region in the complex plane. The number of poles in each region is normalized by the total frame number. Due to this normalization, regions with a value of one correspond to time invariant poles. In Figure 11, we can see no such regions, which indicates that there is no time invariant pole. This result supports assumption (2).

7. CONCLUSION

We have described the problem of speech dereverberation. The contribution of this paper is summarized as follows.

(i) We proposed the joint estimation of the time invariant and time variant subfilters of the inverse filter of an overall acoustic system. It was shown that these subfilters correspond to the inverse filters of a room acoustic system and a speech production system, respectively.

(ii) We developed two distinct algorithms; one uses a criterion based on the SOS of the output while the other is based on the HOS. The SOS-based algorithm improves RASTI by 0.1 even when the observed signals are at most 5-second long. By contrast, the HOS-based algorithm estimates the inverse filter with a RASTI score of nearly one, as long as observed signals of longer than 20 seconds are available.

The main purpose of this paper is to elucidate the theoretical background of the joint estimation based speech dereverberation and the corresponding algorithms and to evaluate their fundamental performance. Thus, we have not

[3] We also conducted an experiment by using real recordings where the room acoustic system might fluctuate and where there was slight background noise. Good dereverberation performance was achieved in this experiment. The result is reported in [27].

investigated practical issues such as computational costs and adaptation to time varying environments. A simple way to cope with these issues would be to employ stochastic gradient learning. An exaustive subjective listening test should also be conducted. Investigating these issues in depth is a subject for future study.

APPENDICES

A. PROOF OF THEOREM 1

By using (2), (5), and (13), we obtain

$$d(n) = \left[\left(1 - A(z,n)\right)\mathbf{G}(z)^T\mathbf{H}(z)\right]s(n). \tag{A.1}$$

Substituting (15) into (A.1) yields

$$\alpha e(n-\beta) = \left[\left(1 - A(z,n)\right)\mathbf{G}(z)^T\mathbf{H}(z)\right]s(n). \tag{A.2}$$

On the other hand, from (9), we have

$$e(n) = \left[1 - B(z,n)\right]s(n) = \left[1 - B(z,n)z^{-\beta}\right]s(n+\beta). \tag{A.3}$$

This equation is equivalent to

$$e(n-\beta) = \left[1 - B(z,n-\beta)z^{-\beta}\right]s(n). \tag{A.4}$$

Relations (A.2) and (A.4) give

$$\begin{aligned}&\left(1 - A(z,n)\right)\mathbf{G}(z)^T\mathbf{H}(z)\\ &\quad = \left(1 - B(z,n-\beta)\right)\alpha z^{-\beta}, \quad 1 \le \forall n \le N.\end{aligned} \tag{A.5}$$

Since both $1 - A(z,n)$ and $1 - B(z,n)$ have no time invariant zero according to (16) and (11), we have

$$\mathbf{G}(z)^T\mathbf{H}(z) = \alpha z^{-\beta}. \tag{A.6}$$

B. DERIVATION OF (26)

In this appendix, we show that $\log|\det\Sigma(\mathbf{d})|$ is invariant with respect to $\{a(k,n)\}_{1\le n\le N,\,1\le k\le P}$ and $\{g_m(k)\}_{1\le m\le M,\,1\le k\le L}$. We here assume that $s(n) = 0$ when $n \le 0$. Hence, relation (B.10), which we derive here, may be an approximation.

Output vector $\mathbf{d}$, defined by (25), is represented by using $\mathbf{y} = [y(N),\ldots,y(1)]^T$ as

$$\mathbf{d} = A\mathbf{y}, \tag{B.1}$$

where A is defined as (B.2):

$$A = \begin{bmatrix}
1 & -a(1,N) & \cdots & \cdots & -a(P,N) & & & & \\
 & 1 & -a(1,N-1) & \cdots & \cdots & -a(P,N-1) & & & \\
 & & \ddots & & & & \ddots & & \\
 & & & 1 & -a(1,P+1) & \cdots & \cdots & -a(P,P+1) & \\
 & & & & 1 & -a(1,P) & \cdots & -a(P-1,P) & \\
 & & & & & \ddots & \ddots & \vdots & \\
 & & & & & & 1 & -a(1,2) & \\
 & & & & & & & 1 &
\end{bmatrix}. \tag{B.2}$$

Relation $\Sigma(\mathbf{d}) = E\{\mathbf{d}\mathbf{d}^T\} = AE\{\mathbf{y}\mathbf{y}^T\}A^T = A\Sigma(\mathbf{y})A^T$ leads to

$$\log\left|\det\Sigma(\mathbf{d})\right| = \log\left|\det\Sigma(\mathbf{y})\right| + 2\log\left|\det A\right|. \tag{B.3}$$

Because the determinant of an upper triangular matrix is the product of its diagonal components, we have $\det A = 1$. Hence, we obtain

$$\log\left|\det\Sigma(\mathbf{d})\right| = \log\left|\det\Sigma(\mathbf{y})\right|. \tag{B.4}$$

$\mathbf{y}$ is related to $\mathbf{s} = [s(N),\ldots,s(1)]^T$ as

$$\mathbf{y} = \sum_{m=1}^{M} G_m\mathbf{x}_m = \left(\sum_{m=1}^{M} G_mH_m\right)\mathbf{s}, \tag{B.5}$$

where $\mathbf{x}_m$, G_m, and H_m are written as

$$\begin{aligned}
\mathbf{x}_m &= \left[x_m(N),\ldots,x_m(1)\right]^T,\\
G_m &= \begin{bmatrix}
g_m(0) & \cdots & g_m(L) & & O \\
 & \ddots & & \ddots & \\
 & & g_m(0) & \cdots & g_m(L) \\
 & & & \ddots & \vdots \\
O & & & & g_m(0)
\end{bmatrix},\\
H_m &= \begin{bmatrix}
h_m(0) & \cdots & h_m(K) & & O \\
 & \ddots & & \ddots & \\
 & & h_m(0) & \cdots & h_m(K) \\
 & & & \ddots & \vdots \\
O & & & & h_m(0)
\end{bmatrix}.
\end{aligned} \tag{B.6}$$

Hence, in a similar way to (B.3), we obtain

$$\begin{aligned}
\log\left|\det\Sigma(\mathbf{y})\right| &= \log\left|\det\Sigma(\mathbf{s})\right| + 2\log\left|\det\left(\sum_{m=1}^{M} G_mH_m\right)\right|\\
&= 2\log\left|\det\left(\sum_{m=1}^{M} G_mH_m\right)\right| + \text{constant}.
\end{aligned} \tag{B.7}$$

Since $\sum_{m=1}^{M} G_mH_m$ is also an upper triangular matrix with diagonal elements of $\sum_{m=1}^{M} h_m(0)g_m(0)$, we have

$$\log\left|\det\left(\sum_{m=1}^{M} G_mH_m\right)\right| = N\log\left(\sum_{m=1}^{M} h_m(0)g_m(0)\right). \tag{B.8}$$

Substituting (18) into (B.8) yields

$$\log\left|\det\left(\sum_{m=1}^{M} G_mH_m\right)\right| = N\log h_1(0) = \text{constant}. \tag{B.9}$$

By using (B.3), (B.7), and (B.9), we can derive

$$\log\det\Sigma(\mathbf{d}) = \text{constant}. \tag{B.10}$$

C. PROOF OF THEOREM 2

By (4) and (12), $d(n)$ is written by using $\{s(n-k)\}_{0\le k\le K+L+P}$ as

$$d(n) = h_1(0)s(n) + L_c\{s(n-k);\ 1 \le k \le K+L+P\}, \tag{C.1}$$

where $L_c\{\cdot\}$ stands for the linear combination. By substituting (8) into (C.1), $d(n)$ is rewritten as

$$d(n) = h_1(0)e(n) + u(n; \mathbf{G}(z), A(z,n)), \tag{C.2}$$

where $u(n)$ is of the form

$$u(n) = L_c\{s(n-k);\ 1 \le k \le K+L+P\}. \tag{C.3}$$

Because $s(n)$ is of the form

$$s(n) = L_c\{e(n), s(n-k);\ 1 \le k \le P\} \tag{C.4}$$

as in (8), $s(n)$ has no components of $\{e(n+k)\}_{k\ge1}$. Therefore, $e(n)$ and $u(n)$ are statistically independent. Then, we have

$$v(d(n)) = h_1(0)^2 v(e(n)) + v(u(n)) \le h_1(0)^2 v(e(n)) \tag{C.5}$$

with equality if and only if

$$v(u(n)) = 0. \tag{C.6}$$

Because the logarithmic function is increasing monotonically, $\sum_{n=1}^{N} \log v(d(n))$ reaches a minimum if and only if

$$v(u(n)) = 0, \quad 1 \le \forall n \le N. \tag{C.7}$$

According to (C.2), condition (C.7) is satisfied if and only if $d(n)$ is equalized with $e(n)$ as

$$d(n) = h_1(0)e(n). \tag{C.8}$$

D. DERIVATION OF (36)

By using the assumption that $d(n)$ is stationary within a single frame and replacing the variance $v(d(n))$ by its sample estimate, the loss function of (33), $\sum_{n=1}^{N} \log v(d(n))$, is estimated by

$$\sum_{i=1}^{T} W \log \langle d(n)^2 \rangle_{n=N_i}^{N_i+W-1} \propto \sum_{i=1}^{T} \log \langle d(n)^2 \rangle_{n=N_i}^{N_i+W-1}. \tag{D.1}$$

The derivative of the right-hand side of (D.1) with respect to $g_m(k)$ is

$$\begin{aligned} &\frac{\partial}{\partial g_m(k)} \sum_{i=1}^{T} \log \langle d(n)^2 \rangle_{n=N_i}^{N_i+W-1} \\ &\quad = \sum_{i=1}^{T} \frac{2}{\langle d(n)^2 \rangle_{n=N_i}^{N_i+W-1}} \left\langle d(n) \frac{\partial d(n)}{\partial g_m(k)} \right\rangle_{n=N_i}^{N_i+W-1}. \end{aligned} \tag{D.2}$$

The derivative of $d(n)$ belonging to the ith frame is

$$\begin{aligned} \frac{\partial d(n)}{\partial g_m(k)} &= \frac{\partial y(n)}{\partial g_m(k)} - \sum_{l=1}^{P} a_i(l) \frac{\partial y(n-l)}{\partial g_m(k)} \\ &= x_m(n-k) - \sum_{l=1}^{P} a_i(l) x_m(n-l-k) \\ &= v_{m,i}(n-k). \end{aligned} \tag{D.3}$$

From (D.2) and (D.3), we have the update equation of (36).

E. PROOF OF THEOREM 3

Let $\{f(k,n)\}_{-\infty\le k\le\infty}$ be the impulse response of the global system $(1 - A(z,n))\mathbf{G}(z)^T\mathbf{H}(z)/(1 - B(z,n))$ at time n. Since $d(n)$ has a non-Gaussian distribution, sequence $\{f(k,n)\}$ has finite nonzero components according to the central limit theorem [24]. Because $d(n)$ is not deterministic, $\{f(k,n)\}$ has at least one nonzero component. Let the first nonzero component of $\{f(k,n)\}$ be $f(\beta_n, n)$. Since the time variant part of the global system $(1 - A(z,n))\mathbf{G}(z)^T\mathbf{H}(z)/(1 - B(z,n))$ has the first tap of weight one, we have

$$\beta_m = \beta_n, \quad f(\beta_m, m) = f(\beta_n, n), \quad \forall m,\ \forall n. \tag{E.1}$$

So we can represent the index and value of the first nonzero component as β and α, respectively. Because variables $\{d(n)\}$ are independent, we obtain the following relation by using Darmois' theorem [25]:

$$f(k,n)f(k-m, n-m) = 0, \quad \forall n,\ \forall k,\ \forall m \ne 0. \tag{E.2}$$

If

$$k = \beta + m, \tag{E.3}$$

we have

$$f(k-m, n-m) = f(\beta, n-m) = \alpha \ne 0. \tag{E.4}$$

Therefore, if $m \neq 0$, we obtain by using (E.2)

$$f(k,n) = f(\beta+m,n) = 0. \tag{E.5}$$

Thus, $\{f(k,n)\}$ has only one nonzero component $f(\beta,n) = \alpha$. Since $d(n)$ is represented as

$$d(n) = \left[\frac{(1-A(z,n))\mathbf{G}(z)^T\mathbf{H}(z)}{1-B(z,n)}\right]e(n), \tag{E.6}$$

$d(n)$ is equalized with $e(n)$ up to constant scale α and delay β.

F. DERIVATION OF (40)

Mutual information $\mathcal{I}(d(1),\ldots,d(N))$ is defined as

$$\mathcal{I}(d(1),\ldots,d(N)) = \sum_{n=1}^{N}\mathcal{H}(d(n)) - \mathcal{H}(\mathbf{d}), \tag{F.1}$$

where $\mathcal{H}(\boldsymbol{\xi})$ represents the differential entropy of (multivariate) random variable $\boldsymbol{\xi}$. From (B.1), we have

$$\mathcal{H}(\mathbf{d}) = \mathcal{H}(\mathbf{y}) + \log|\det A|. \tag{F.2}$$

Because of (B.3), we also have

$$\log|\det A| = \frac{1}{2}\left(\log|\det\Sigma(\mathbf{d})| - \log|\det\Sigma(\mathbf{y})|\right). \tag{F.3}$$

Substituting (F.2) and (F.3) into (F.1) gives

$$\begin{aligned}
&\mathcal{I}(d(1),\ldots,d(N))\\
&\quad= \sum_{n=1}^{N}\mathcal{H}(d(n)) - \frac{1}{2}\log|\det\Sigma(\mathbf{d})|\\
&\qquad+ \frac{1}{2}\log|\det\Sigma(\mathbf{y})| - \mathcal{H}(\mathbf{y})\\
&\quad= -\sum_{n=1}^{N}\left(\frac{1}{2}\log v(d(n)) - \mathcal{H}(d(n))\right)\\
&\qquad+ \frac{1}{2}\left(\sum_{n=1}^{N}\log v(d(n)) - \log|\det\Sigma(\mathbf{d})|\right)\\
&\qquad+ \frac{1}{2}\log|\det\Sigma(\mathbf{y})| - \mathcal{H}(\mathbf{y}).
\end{aligned} \tag{F.4}$$

Now, the negentropy of n-dimensional random variable $\boldsymbol{\xi}$ is defined as

$$\begin{aligned}
\mathcal{J}(\boldsymbol{\xi}) &= \mathcal{H}(\boldsymbol{\xi}^{\text{gauss}}) - \mathcal{H}(\boldsymbol{\xi})\\
&= \frac{1}{2}\log|\det\Sigma(\boldsymbol{\xi}^{\text{gauss}})| + \frac{n}{2}(1+\log 2\pi) - \mathcal{H}(\boldsymbol{\xi}),
\end{aligned} \tag{F.5}$$

where $\boldsymbol{\xi}^{\text{gauss}}$ is a Gaussian random variable with the same covariance matrix as that of $\boldsymbol{\xi}$. By using (20) and (F.5), (F.4) is rewritten as

$$\begin{aligned}
&\mathcal{I}(d(1),\ldots,d(N))\\
&\quad= -\sum_{n=1}^{N}\mathcal{J}(d(n)) + \mathcal{J}(\mathbf{y}) + \mathcal{K}(d(1),\ldots,d(N)).
\end{aligned} \tag{F.6}$$

Furthermore, since $\mathbf{y}$ is related to $\mathbf{s}$ by an $N\times N$ regular linear transformation according to (B.5), and the negentropy is conserved by such linear transformation, we obtain

$$\mathcal{J}(\mathbf{y}) = \text{constant}. \tag{F.7}$$

From (F.6) and (F.7), we finally reach (42).

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Upper Saddle River, NJ, USA, 1983.

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

[3] K. Furuya and Y. Kaneda, "Two-channel blind deconvolution of nonminimum phase FIR systems," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E80-A, no. 5, pp. 804–808, 1997.

[4] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[5] T. Hikichi, M. Delcroix, and M. Miyoshi, "Blind dereverberation based on estimates of signal transmission channels without precise information on channel order," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, pp. 1069–1072, Philadelphia, Pa, USA, March 2005.

[6] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Transactions Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 430–440, 2007.

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

[8] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *IEEE Interntional Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 6, pp. 3701–3704, Salt Lake, Utah, USA, May 2001.

[9] B. W. Gillespie and L. E. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," in *IEEE International Conference on Accoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 676–679, Hong Kong, April 2003.

[10] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 99–102, Kyotp, Japan, September 2003.

[11] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Transactions, Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 80–95, 2007.

[12] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Efficient blind dereverberation framework for automatic speech recognition," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 3145–3148, Lisbon, Portugal, September 2005.

[13] P. S. Spencer and P. J. W. Rayner, "Separation of stationary and time-varying systems and its application to the restoration of gramophone recordings," in *IEEE International Symposium on Circuits and Systems (ISCAS '89)*, vol. 1, pp. 292–295, Portland, Ore, USA, May 1989.

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

[15] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems(channels)," *IEEE Transactions on Information Theory*, vol. 36, no. 2, pp. 312–321, 1990.

[16] K. Abed-Meraim, E. Moulines, and P. Loubaton, "Prediction error method for second-order blind identification," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 694–705, 1997.

[17] B. Theobald, S. Cox, G. Cawley, and B. Milner, "Fast method of channel equalisation for speech signals and its implementation on a DSP," *Electronics Letters*, vol. 35, no. 16, pp. 1309–1311, 1999.

[18] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[19] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[20] Acoustical Society of Japan, "ASJ Continuous Speech Corpus," http://www.mibel.cs.tsukuba.ac.jp/jnas/instruct.html.

[21] H. Kuttruff, *Room Acoustics*, Elsevier Applied Science, London, UK, 1991.

[22] W. B. Kleijn and K. K. Paliwal, Eds., *Speech Coding and Synthesis*, Elsevier Science, Amsterdam, The Netherlands, 1995.

[23] A. Gorokhov and P. Loubaton, "Blind identification of MIMO-FIR systems: a generalized linear prediction approach," *Signal Processing*, vol. 73, no. 1-2, pp. 105–124, 1999.

[24] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes*, Springer, New York, NY, USA, 1987.

[25] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[26] A. Hyvärinen, J. Karhumen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, New York, NY, USA, 2001.

[27] T. Yoshioka, T. Hikichi, M. Miyoshi, and H. G. Okuno, "Robust decomposition of inverse filter of channel and prediction error filter of speech signal for dereverberation," in *Proceedings of the 14th European Signal Processing Conference (EUSIPCO '06)*, Florence, Italy, 2006.

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

Takuya Yoshioka received the M.S. of Informatics degree from Kyoto University, Kyoto, Japan, in 2006. He is currently with the Signal Processing Group of NTT Communication Science Laboratories. His research interests are in speech and audio signal processing and statistical learning.

Takafumi Hikichi was born in Nagoya, in 1970. He received his B.S. and M.S. of electrical engineering degrees from Nagoya University in 1993 and 1995, respectively. In 1995, he joined the Basic Research Laboratories of NTT. He is currently working at the Signal Processing Research Group of the Communication Science Laboratories, NTT. He is a Visiting Associate Professor of the Graduate School of Information Science, Nagoya University. His research interests include physical modeling of musical instruments, room acoustic modeling, and signal processing for speech enhancement and dereverberation. He received the 2000 Kiyoshi-Awaya Incentive Awards, and the 2006 Satoh Paper Awards from the ASJ. He is a Member of IEEE, ASA, ASJ, and IEICE.

Masato Miyoshi received his M.E. degree from Doshisha University in Kyoto in 1983. 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, the Media Information Laboratory, NTT Communication Science Labs. He is also a Visiting Associate Professor of the Graduate School of Information Science and Technology, Hokkaido University. He was honored to receive the 1988 IEEE senior awards, the 1989 ASJ Kiyoshi-Awaya incentive awards, the 1990 and 2006 ASJ Sato Paper awards, and the 2005 IEICE Paper awards, respectively. He also received his Ph.D. degree from Doshisha University in 1991. He is a Member of IEICE, ASJ, AES, and a Senior Member of IEEE.