14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

ROBUST DECOMPOSITION OF INVERSE FILTER OF CHANNEL AND PREDICTION ERROR FILTER OF SPEECH SIGNAL FOR DEREVERBERATION

Takuya Yoshioka,[†‡] *Takafumi Hikichi,*[†] *Masato Miyoshi,*[†] *and Hiroshi G. Okuno*[‡]

[†]NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237, Japan
[‡]Graduate School of Informatics, Kyoto University
Yoshida-hommachi, Sakyo-ku, Kyoto, 606-8501, Japan
email: {takuya, hikichi, miyo}@cslab.kecl.ntt.co.jp, okuno@i.kyoto-u.ac.jp

ABSTRACT

This paper estimates the inverse filter of a signal transmission channel of a room driven by a speech signal. Speech signals are often modeled as piecewise stationary autoregressive (AR) processes. The most fundamental issue is how to estimate a channel's inverse filter separately from the inverse filter of the speech generating AR system, or the prediction error filter (PEF). We first point out that by jointly estimating the channel's inverse filter and the PEF, the channel's inverse is identifiable due to the time varying nature of the PEF. Then, we develop an algorithm that achieves this joint estimation. The notable property of the proposed method is its robustness against deviation from the linear convolutive model of an observed signal caused by, for example, observation noise. Experimental results with simulated and real recorded reverberant signals showed the effectiveness of the proposed method.

1. INTRODUCTION

Room reverberation degrades speech intelligibility or corrupts characteristics inherent in speech. Therefore, cancelling the effect of the reverberation is indispensable for a variety of speech processing applications such as hands-free telephony or automatic speech recognition. Since only a reverberant speech signal is available in many practical situations, the dereverberation should be based on blind processing, which indicates a form of processing that operates solely with a microphone signal.

Let a source signal be represented by $s(n)$, and K-tap impulse responses from the source to M microphones by $\{h_1(k)\}_{k=0}^{K},\cdots,\{h_M(k)\}_{k=0}^{K}$. Microphone signals $x_1(n),\cdots,x_M(n)$ summarized in vector $\boldsymbol{x}(n) = [x_1(n),\cdots,x_M(n)]^*$, where superscript * indicates the transposition of a vector or a matrix, can be modeled as

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

where $\boldsymbol{h}(k) = [h_1(k),\cdots,h_M(k)]^*$. This can be written by using the vector, $\boldsymbol{H}(z) = \sum_{k=0}^{K}\boldsymbol{h}(k)z^{-k}$, of the room transfer functions (RTFs) as

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

where $[z^{-1}]$ represents a backward shift operator. Then, we may formulate the dereverberation as

$$y(n) = \sum_{k=0}^{L} \boldsymbol{g}(k)^*\boldsymbol{x}(n-k), \text{ or } y(n) = [\boldsymbol{G}(z)^*]\boldsymbol{x}(n), \tag{3}$$

where $\boldsymbol{g}(k) = [g_1(k),\cdots,g_M(k)]^*$ is a vector of the coefficients of the L-tap inverse filter set, $\boldsymbol{G}(z) = \sum_{k=0}^{L}\boldsymbol{g}(k)z^{-k}$, of the RTFs $\boldsymbol{H}(z)$. In the context of blind dereverberation, we want to set up the transfer function vector $\boldsymbol{G}(z)$ so that it can provide the inverse of $\boldsymbol{H}(z)$ up to a constant scale and delay as

$$\boldsymbol{G}(z)^*\boldsymbol{H}(z) = \alpha z^{-\beta} \tag{4}$$

without any specific knowledge about $s(n)$ or $\boldsymbol{H}(z)$.

And now, a speech signal can be modeled as a piecewise stationary autoregressive (AR) process [1]. In this model, a signal in the ith time frame is described as

$$s(n) = \sum_{k=1}^{P} b_i(k)s(n-k) + e_i(n), \text{ or } s(n) = \Big[\frac{1}{1-B_i(z)}\Big]e_i(n) \tag{5}$$

where $1-B_i(z) = 1-\sum_{k=1}^{P} b_i(k)z^{-k}$ and $e_i(n)$ denote a prediction error filter (PEF) and an innovation, respectively. Thus, $\boldsymbol{x}(n)$ is the output of the system, which is a cascade consisting of $1/(1-B_i(z))$ and $\boldsymbol{H}(z)$ excited by $e_i(n)$. The most fundamental problem is then how to identify the inverse of $\boldsymbol{H}(z)$ separately from the PEF.

Several methods have been proposed for estimating the inverse of $\boldsymbol{H}(z)$ and the PEF separately [2, 3, 4]. However, these methods are sensitive to deviation from the linear convolutive model of Eq. (2) caused by indeterminacy of the order K or observation noise. This characteristic makes it difficult to apply these methods to a real environment. More investigation is required to overcome this difficulty.

Another previously considered approach relies on the observation that the PEF $1-B_i(z)$ can be approximately computed by applying linear prediction [1] directly to the observed signals $\boldsymbol{x}(n)$. This is probably because short-term correlations in reverberant speech signals mainly reflect the speech characteristics whereas long-term correlations reflect those of the reverberation. Based on this observation, methods have been proposed for estimating the PEF in every short time frame followed by inverse filter set estimation [5, 6]. Although this class of method is robust even when the signals $\boldsymbol{x}(n)$ do not strictly obey the model given by Eq. (2), the performance remains insufficient. The underlying reason for this drawback is the mutual dependency between the estimate of the RTF inverse filter set and estimates of the PEFs; precise PEF estimation requires a good estimate of the inverse filter set and vice versa.

In this paper, we approach this problem by jointly estimating $\boldsymbol{G}(z)$ and $1-A_i(z)$, by which we denote an estimate of $1-B_i(z)$. The time varying nature of the PEFs means we can

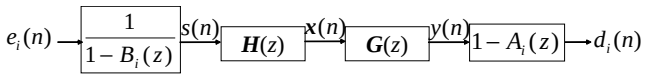

Figure 1: Schematic diagram of overall system.

uniquely identify $\boldsymbol{G}(z)$ satisfying Eq. (4) and $1-A_i(z)$ equal to $1-B_i(z)$. This becomes possible by equalizing the signal $d_i(n)$ that is produced by filtering $\boldsymbol{x}(n)$ through $\boldsymbol{G}(z)$ and $1-A_i(z)$ (see Fig. 1) with the innovation $e_i(n)$. Since $e_i(n)$ is unavailable in reality, the estimation of $\boldsymbol{G}(z)$ and $1-A_i(z)$ is achieved instead by maximizing the loss function defined as the mutual information between the samples of the sequence $\{d_i(n)\}_{i,n}$ based on the mutual independence property of the innovations. Importantly, because a small deviation from the model of Eq. (2) causes only a slight change in the loss function, the estimation is expected to be robust to such a deviation. Direct optimization is very complicated, so we introduce some approximations to derive a simple estimation algorithm. The proposed method was tested on simulated reverberant speech as well as real recordings. The results showed that the proposed method improved the speech intelligibility and reduced the spectral distortion from the clean speech. Note that, in constrast to the previous dereverberation method, which utilizes the time varying nature of the PEFs [7], the proposed method can estimate the (delayed) inverse of a nonminimum phase channel.

2. SYSTEM DECOMPOSITION BASED ON TIME VARYING NATURE

Let us consider the system illustrated in Fig. 1. This diagram shows the overall system for producing an innovation estimate. Source signal $s(n)$ and observed signals $\boldsymbol{x}(n)$ are modeled as Eqs. (5) and (2), respectively. Then, the reverberant signal is passed into inverse filter set $\boldsymbol{G}(z)$ to generate inverse filtered signal $y(n)$ as Eq. (3). The inverse filtered signal is further segmented into T short time frames. Finally, the inverse filtered signal $y_i(n)$ in the ith time frame is filtered by $1-A_i(z)$ to produce the estimate, $d_i(n)$, of the innovation $e_i(n)$ as

$$d_i(n) = y_i(n) - \sum_{k=1}^{P} a_i(k) y_i(n-k), \text{ or } d_i(n) = [1-A_i(z)]y_i(n). \tag{6}$$

We consider the case where both $s(n)$ and $y(n)$ are segmented by a W-sample rectangular window. The relationship between $s(n)$ and $s_i(n)$ is as follows:

$$s_i(n) = s((i-1)W+n). \tag{7}$$

The same applies to relationship between $y(n)$ and $y_i(n)$.

We make the following assumptions:

P1) There is no zero common to all of the PEFs $1-B_1(z),\cdots,1-B_T(z)$.

P2) The innovation sequence $\{e_i(n)\}_{n=1}^{W}$ in the ith time frame consists of zero-mean i.i.d. random variables. The distribution of $e_i(n)$ is symmetric and supergaussian. Innovations in different time frames are also mutually independent, but not necessarily identically distributed.

Then, we can present the following theorem.

Theorem 1. *If $d_i(n) = \alpha e_{i-\gamma}(n)$, and there is no zero common to $1-A_1(z),\cdots,1-A_T(z)$, then $\boldsymbol{G}(z)^*\boldsymbol{H}(z) = \alpha z^{-W\gamma}$.*

Proof. $d_i(n) = \alpha e_{i-\gamma}(n)$ indicates

$$\frac{1-A_i(z)}{1-B_{i-\gamma}(z)}\boldsymbol{G}(z)^*\boldsymbol{H}(z) = \alpha z^{-W\gamma}. \tag{8}$$

Then, we have

$$(1-A_i(z))\boldsymbol{G}(z)^*\boldsymbol{H}(z) = (1-B_{i-\gamma}(z))\alpha z^{-W\gamma}. \tag{9}$$

The no common zero condition for $1-A_i(z)$ and $1-B_i(z)$ among different i's leads directly to the theorem. □

Remark. The constant delay β in Eq. (4) is here limited to a multiple of the window length W.

As a consequence of Theorem 1, the inverse of $\boldsymbol{H}(z)$ can be uniquely identified from $\boldsymbol{G}(z)$ and $1-A_i(z)$ computed so that $d_i(n)$ is equalized with $\alpha e_{i-\gamma}(n)$. If there is the term $\tilde{A}(z)$ common to all of the computed $1-A_1(z),\cdots,1-A_T(z)$, we just have to replace $\boldsymbol{G}$ by $\boldsymbol{G}\tilde{A}(z)$. Thus, jointly estimating $\boldsymbol{G}(z)$ and $1-A_i(z)$ enables us to estimate the inverse filter set separately from the PEFs. The most noteworthy consequence is that the joint estimation is vital for highly accurate inverse filter estimation since both $\boldsymbol{G}(z)$ and $1-A_i(z)$ are mutually dependent, and contribute to the production of $d_i(n)$ cooperatively.

3. BLIND PROCESSING BASED ALGORITHM

3.1 Loss function

According to the above discussion, we need to estimate $\boldsymbol{G}(z)$ and $1-A_i(z)$ so that $d_i(n)$ is equalized with $e_i(n)$ up to a constant scale and delay. However, since innovation $e_i(n)$ is unavailable, we have to develop a criterion computed solely by using $d_i(n)$.

Based on condition P2), it would be natural to estimate $\boldsymbol{G}(z)$ and $1-A_i(z)$ so that the inter-sample dependence of $\{d_i(n)\}_{i,n}$ is minimized. This is mathematically formulated as

$$\begin{aligned}\{\hat{\boldsymbol{g}},\hat{\boldsymbol{a}}\} = \underset{\boldsymbol{g},\boldsymbol{a}}{\operatorname{argmin}} & \mathcal{I}(d_1(1),\cdots,d_1(W),\cdots,d_T(1),\cdots,d_T(W)) \\ & \text{subject to } ||\boldsymbol{g}|| = 1 \text{ and} \\ & \quad 1-A_i(z) \text{ is minimum phase,}\end{aligned} \tag{10}$$

where $\boldsymbol{g} = [\boldsymbol{g}_1{}^*,\cdots,\boldsymbol{g}_M{}^*]^*$, $\boldsymbol{g}_m = [g_m(0),\cdots,g_m(L)]^*$, $\boldsymbol{a} = [\boldsymbol{a}_1{}^*,\cdots,\boldsymbol{a}_T{}^*]^*$, $\boldsymbol{a}_i = [a_i(1),\cdots,a_i(P)]^*$, and $\mathcal{I}(\xi_1,\cdots,\xi_n)$ is the mutual information between random variables $\xi_1,\cdots,\xi_n$. The first constraint of Eq. (10) determines the constant scale arbitrarily. The second corresponds to the minimum phase property of PEF $1-B_i(z)$. In the rest of Sect. 3.1, we give a more specific representation of the loss function $\mathcal{I}(d_1(1),\cdots,d_T(W))$. Then, the estimation algorithm is explained in Sect. 3.2 to 3.4.

The mutual information in Eq. (10) is defined as

$$\mathcal{I}(d_1(1),\cdots,d_T(W)) = \sum_{i=1}^{T}\sum_{n=1}^{W}\mathcal{H}(d_i(n)) - \mathcal{H}(\boldsymbol{d}), \tag{11}$$

where $\boldsymbol{d} = [d_T(W),\cdots,d_1(1)]^*$ and $\mathcal{H}(\boldsymbol{\xi})$ denotes the differential entropy of random variable $\boldsymbol{\xi}$. The output signal vector $\boldsymbol{d}$ is represented with respect to the inverse filtered signal

vector $\boldsymbol{y} = [y_T(W), \cdots, y_1(1)]^*$ as

$$\boldsymbol{d} = A\boldsymbol{y}, \tag{12}$$

where

$$A = \begin{bmatrix} A_T & & O \\ & \ddots & \\ O & & A_1 \end{bmatrix} \tag{13}$$

$$A_i = \begin{bmatrix} 1 & -a_i(1) & \cdots & -a_i(P) & & O \\ & \ddots & \ddots & & \ddots & \\ & & 1 & -a_i(1) & \cdots & -a_i(P) \\ & & & \ddots & \ddots & \vdots \\ & & & & \ddots & -a_i(1) \\ O & & & & & 1 \end{bmatrix}. \tag{14}$$

Hence, the differential entropy $\mathcal{H}(\boldsymbol{d})$ can be written as

$$\mathcal{H}(\boldsymbol{d}) = \mathcal{H}(\boldsymbol{y}) + \log\det A. \tag{15}$$

Let us denote the covariance matrix of multivariate random variable $\boldsymbol{\xi}$ by $\Sigma(\boldsymbol{\xi})$. Then, equation $\Sigma(\boldsymbol{d}) = E\{\boldsymbol{d}\boldsymbol{d}^*\} = AE\{\boldsymbol{y}\boldsymbol{y}^*\}A^* = A\Sigma(\boldsymbol{y})A^*$ leads to

$$\log\det A = \frac{1}{2}\big(\log\det\Sigma(\boldsymbol{d}) - \log\det\Sigma(\boldsymbol{y})\big). \tag{16}$$

Substituting Eqs. (15) and (16) into Eq. (11) yields

$$\begin{aligned}\mathcal{I}(d_1(1), \cdots, d_T(W)) = &-\sum_{i=1}^{T}\sum_{n=1}^{W}\mathcal{J}(d_i(n)) + \mathcal{J}(\boldsymbol{y}) \\ &+ \mathcal{K}(d_1(1), \cdots, d_T(W)).\end{aligned} \tag{17}$$

In Eq. (17), $\mathcal{J}(\boldsymbol{\xi})$ denotes negentropy [8], which is a measure of the nongaussianity of random variable $\boldsymbol{\xi}$. $\mathcal{K}(\xi_1, \cdots, \xi_n)$ is defined as

$$\mathcal{K}(\xi_1, \cdots, \xi_n) = \frac{1}{2}\Big(\sum_{i=1}^{n}\log\upsilon(\xi_i) - \log\det\Sigma([\xi_1, \cdots, \xi_n]^*)\Big), \tag{18}$$

where $\upsilon(\xi_1), \cdots, \upsilon(\xi_n)$ represent the variances of random variables $\xi_1, \cdots, \xi_n$, respectively. This is a measure of the correlatedness of $\xi_1, \cdots, \xi_n$ [9]. If we put $\boldsymbol{s} = [s_T(W), \cdots, s_1(1)]^*$, we can readily obtain

$$\mathcal{J}(\boldsymbol{y}) = \mathcal{J}(\boldsymbol{s}) = \text{constant} \tag{19}$$

since $\boldsymbol{y}$ is a linear transformation of $\boldsymbol{s}$. By using Eqs. (17) and (19), Eq. (10) reduces to

$$\begin{aligned}\{\hat{\boldsymbol{g}}, \hat{\boldsymbol{a}}\} = \underset{\boldsymbol{g},\boldsymbol{a}}{\operatorname{argmin}}\Big(&-\sum_{i=1}^{T}\sum_{n=1}^{W}\mathcal{J}(d_i(n)) + \mathcal{K}(d_1(1), \cdots, d_T(W))\Big) \\ &\text{subject to } ||\boldsymbol{g}|| = 1 \text{ and} \\ &1 - A_i(z) \text{ is minimum phase.}\end{aligned} \tag{20}$$

In this way, the minimization of $\mathcal{I}(d_1(1), \cdots, d_T(W))$ represented by Eq. (10) is equivalent to the minimization of the loss function composed of the negentropy, $\mathcal{J}(d_i(n))$, of $d_i(n)$ and the correlatedness, $\mathcal{K}(d_1(1), \cdots, d_T(W))$, of $d_1(1), \cdots, d_T(W)$ as in Eq. (20).

3.2 Estimation by alternating variables method

We solve the optimization problem of Eq. (20) by employing an alternating variables method. Let $\hat{\boldsymbol{a}}^{(t)}$ and $\hat{\boldsymbol{g}}^{(t)}$ denote the estimates of $\boldsymbol{a}$ and $\boldsymbol{g}$ obtained after the tth iteration, respectively. The estimates of the $(t+1)$th iteration are then computed by solving the following optimization problems:

$$\begin{aligned}\hat{\boldsymbol{a}}^{(t+1)} = \underset{\boldsymbol{a}}{\operatorname{argmin}}\Big(&-\sum_{i=1}^{T}\sum_{n=1}^{W}\mathcal{J}(d_i(n)) + \mathcal{K}(d_1(1), \cdots, d_T(W))\Big) \\ &\text{subject to } \boldsymbol{g} = \hat{\boldsymbol{g}}^{(t)} \text{ and } 1 - A_i(z) \text{ is minimum phase}\end{aligned} \tag{21}$$

and

$$\begin{aligned}\hat{\boldsymbol{g}}^{(t+1)} = \underset{\boldsymbol{g}}{\operatorname{argmin}}\Big(&-\sum_{i=1}^{T}\sum_{n=1}^{W}\mathcal{J}(d_i(n)) + \mathcal{K}(d_1(1), \cdots, d_T(W))\Big) \\ &\text{subject to } \boldsymbol{a} = \hat{\boldsymbol{a}}^{(t+1)} \text{ and } ||\boldsymbol{g}|| = 1.\end{aligned} \tag{22}$$

The algorithms used to accomplish Eqs. (21) and (22) are given in Sect. 3.3 and 3.4, respectively.

It should be noted that if we set $\hat{\boldsymbol{g}}^{(0)} = [1, 0, \cdots, 0]^*$ and accept $\hat{\boldsymbol{g}}^{(1)}$ as a final estimate of $\boldsymbol{g}$, the proposed method becomes similar to conventional ones [5, 6] in that the PEFs are estimated directly from the observed signals. By updating the estimates of $\boldsymbol{g}$ and $\boldsymbol{a}$ iteratively, however, the proposed method becomes capable of the joint estimation of $\boldsymbol{g}$ and $\boldsymbol{a}$, which we have pointed out as being vital for highly accurate inverse filter estimation.

3.3 Estimation of prediction error filters

In solving the optimization problem of Eq. (21), we only minimize the second term, $\mathcal{K}(d_1(1), \cdots, d_T(W))$, of the loss function. This is because minimizing the first term $-\sum_{i=1}^{T}\sum_{n=1}^{W}\mathcal{J}(d_i(n))$ might make $1 - A_i(z)$ nonminimum phase since the negentropy $\mathcal{J}(d_i(n))$ is related to the higher order statistics of $d_i(n)$.

From Eq. (18), $\mathcal{K}(d_1(1), \cdots, d_T(W))$ is represented as

$$\mathcal{K}(d_1(1), \cdots, d_T(W)) = \frac{1}{2}\Big(\sum_{i=1}^{T}\sum_{n=1}^{W}\log\upsilon(d_i(n)) - \log\det\Sigma(\boldsymbol{d})\Big). \tag{23}$$

Because the determinant of an upper triangular matrix is the product of its diagonal components, we have $\log\det A = 0$ from Eqs. (13) and (14). Substituting this relation into Eq. (16) leads to

$$\log\det\Sigma(\boldsymbol{d}) = \log\det\Sigma(\boldsymbol{y}) = \text{constant}. \tag{24}$$

Thus, the minimization of $\mathcal{K}(d_1(1), \cdots, d_T(W))$ is equivalent to the minimization of the variance of $d_i(n)$. This minimization is accomplished by applying the linear prediction to the inverse filtered and segmented signal $y_i(n)$. It should be noted that the linear prediction guarantees $1 - A_i(z)$ to be minimum phase [1].

3.4 Estimation of inverse filter set

When solving the optimization problem of Eq. (22), the term $\mathcal{K}(d_1(1), \cdots, d_T(W))$ is negligible compared with $\sum_{i=1}^{T}\sum_{n=1}^{W}J(d_i(n))$ since $\mathcal{K}(d_1(1), \cdot, d_T(W))$ is minimized in

the previous estimation of $\boldsymbol{a}$. Based on P2) and by using the Gram-Charlier expansion, the negentropy $\mathcal{J}(d_i(n))$ can be approximated by [8]

$$\mathcal{J}(d_i(n)) \simeq \frac{\kappa_4(d_i(n))^2}{48\upsilon(d_i(n))^4} + \text{constant}, \tag{25}$$

where $\kappa_4(\xi)$ denotes the kurtosis of random variable ξ. Because the innovation of a speech signal is supergaussian as in P2), we solve the following optimization problem:

$$\boldsymbol{g}^{(t+1)} = \underset{\boldsymbol{g}}{\text{argmax}}\, Q \quad \text{subject to } \boldsymbol{a} = \hat{\boldsymbol{a}}^{(t)} \text{ and } ||\boldsymbol{g}|| = 1, \tag{26}$$

where

$$Q = \frac{1}{W}\sum_{i=1}^{T}\sum_{n=1}^{W}\frac{\kappa_4(d_i(n))}{\upsilon(d_i(n))^2}. \tag{27}$$

Q is maximized by using the gradient method. Based on the stationarity assumption in each time frame, the normalized kurtosis $\kappa_4(d_i(n))/\upsilon(d_i(n))^2$ is approximated by its sample estimate $\langle d_i(n)^4\rangle/\langle d_i(n)^2\rangle^2 - 3$, where $\langle\cdot\rangle$ denotes an averaging operator. By calculating the derivative of $Q \simeq \sum_{i=1}^{T}\langle d_i(n)^4\rangle/\langle d_i(n)^2\rangle^2 - 3$ with respect to $\boldsymbol{g}$, we have the following update equation:

$$\boldsymbol{g}' = \boldsymbol{g}^{\langle u\rangle} + \eta\nabla Q_{\boldsymbol{g}}(\boldsymbol{g}^{\langle u\rangle}), \tag{28}$$

$$\boldsymbol{g}^{\langle u+1\rangle} = \frac{\boldsymbol{g}'}{||\boldsymbol{g}'||} \tag{29}$$

where

$$\nabla Q_{\boldsymbol{g}} = \left[\frac{\partial Q}{\partial g_1(0)},\cdots,\frac{\partial Q}{\partial g_1(L)},\cdots,\frac{\partial Q}{\partial g_M(0)},\cdots,\frac{\partial Q}{\partial g_M(L)}\right]^* \tag{30}$$

$$\frac{\partial Q}{\partial g_m(k)} = \sum_{i=1}^{T}\frac{4}{\langle d_i(n)^2\rangle^4}\big(\langle d_i(n)^3 v_{mi}(n-k)\rangle\langle d_i(n)^2\rangle^2 - \langle d_i(n)^4\rangle\langle d_i(n)^2\rangle\langle d_i(n)v_{mi}(n-k)\rangle\big) \tag{31}$$

$$v_{mi}(n) = x_{mi}(n) - \sum_{k=1}^{P} a_i(k)x_{mi}(n-k) \tag{32}$$

$$x_{mi}(n) = x_m((i-1)W+n), \tag{33}$$

u is an iteration number, and η is a step size. Note that the update procedure defined by Eq. (28) to (33) is different from the procedure derived in [6] in that the former explicitly employs framewise kurtosis normalization. The framewise normalization would be important since the variance of the innovation $e_i(n)$ changes frame by frame.

4. EXPERIMENTAL RESULTS

We conducted experiments to evaluate the dereverberation performance of the proposed method. Male and female speech signals of Japanese sentences taken from the ASJ-JNAS database were used as the clean speech signals. These speech signals satisfied the assumptions P1) and P2) well. The signals were sampled at 8 kHz and quantized with 16-bit resolution. We first show results obtained by using simulated reverberant speech signals in order to assess the dereverberation performance of the proposed method. Then, we also present results obtained with reverberant speech signals recorded in a real room.

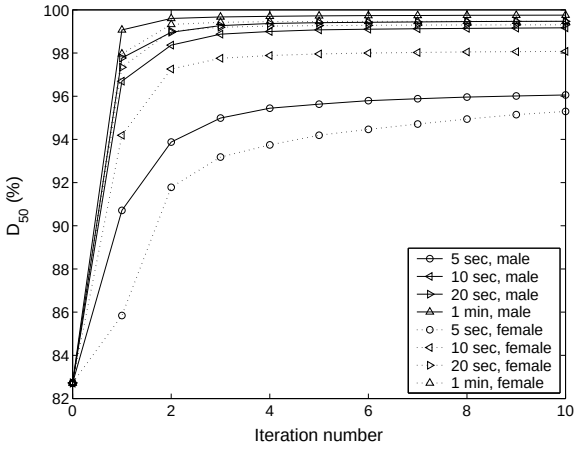

Figure 2: D_{50} as a function of the number of iterations.

4.1 Simulated reverberant speech

The microphone signals were simulated by convolving the clean speech signals with impulse responses measured in a reverberation room. The room size was $4.45 \times 3.55 \times 2.5$ m. The distance between the loudspeaker and the microphones was about 3.2 m. The reverberation time was around 0.5 sec. The microphone signal was prewhitened before it entered the proposed algorithm in order to stabilize the gradient algorithm of Eqs. (28) and (29).

The following parameter settings were used: $M = 4$, $L = 1000$, $W = 200$, $P = 16$. The variables $\boldsymbol{g}$ and $\boldsymbol{a}$ were alternated 10 times. The estimate of $\boldsymbol{g}^{(t+1)}$ was updated 20 times by using Eqs. (28) and (29). The initial estimate of the inverse filters was set as

$$g_m(k) = \begin{cases} 1/M & \text{if } k = 200 \\ 0 & \text{otherwise} \end{cases}. \tag{34}$$

Although we used $M = 4$ microphones here, the proposed method is potentially applicable to $M = 1$. The main advantage of using multiple microphones rather than just a single microphone is the ability to estimate an inverse filter with a small number of observed signal.

The dereverberation performance was evaluated by using D_{50} [10], which is a measure related to speech intelligibility. The measure D_{50} is defined as

$$D_{50} = \frac{\int_0^{50\text{ msec.}} f(t)^2 dt}{\int_0^{\infty} f(t)^2 dt} \times 100\ (\%), \tag{35}$$

where $\{f(t); t \geq 0\}$ denotes an arbitrary impulse response.

Figure 2 shows the dereverberation performance against the number of iterations. The dereverberation performance improved with the number of iterations under all conditions. The iteration was effective especially when a small amount of speech was available.

4.2 Recorded reverberant speech

Finally, we show results obtained by applying the proposed method to reverberant speech recorded in the same room with a slightly different speaker and microphone configuration. The reverberation time was around 0.45 sec. The (noisy)

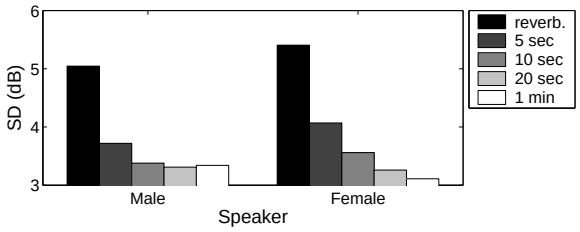

Figure 3: Average SDs for reverberant and dereverberated speech signals.

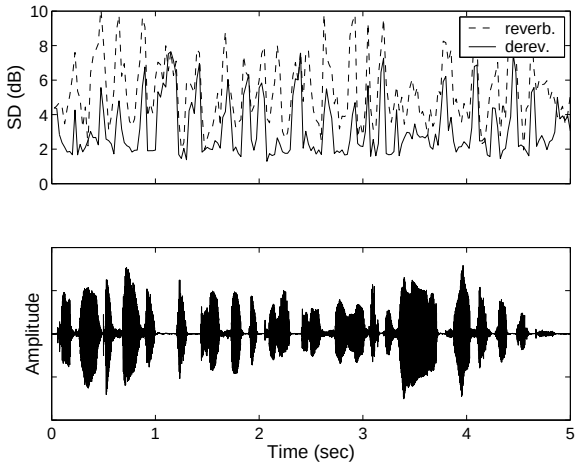

Figure 4: (Top panel) SDs as a function of time for female speech. The inverse filter set was estimated by using 20-sec speech. (Bottom panel) Speech waveform of corresponding portion.

speech level was 25 to 35 dB above the background noise level. The parameters were set as in the last section.

The dereverberation performance was measured in terms of the spectral distortion (SD) between $1/(1-B_i(z))$ and $1/(1-A_i(z))$ defined as

$$SD = \sqrt{\frac{1}{F}\sum_{f=0}^{F-1}\left(20\log|P_A(f)| - 20\log|P_B(f)|\right)^2}\ (\mathrm{dB}), \quad (36)$$

where $P_A(f) = 1/(1-A_i(e^{\frac{j\pi f}{F}}))$, $P_B(f) = 1/(1-B_i(e^{\frac{j\pi f}{F}}))$, and F is the number of frequency bins. F was set at 256.

Figure 3 shows the SDs averaged over all the time frames. It can be seen that the SDs were improved by the proposed method. More importantly, the SDs were small during the voiced portions of the speech (see Fig. 4). This indicates that the proposed method recovered the speech characteristics well. Figure 5 shows example speech spectrograms. The inverse filtered speech seems to be dereverberated well.

5. CONCLUSIONS

We have described a novel blind dereverberation method for speech signals. The method jointly estimates an inverse filter of a signal transmission channel and PEFs of the speech signals in every short time frame. A simple iterative estimation procedure was derived. The proposed method achieved good dereverberation for simulated reverberant speech as well as real recorded speech.

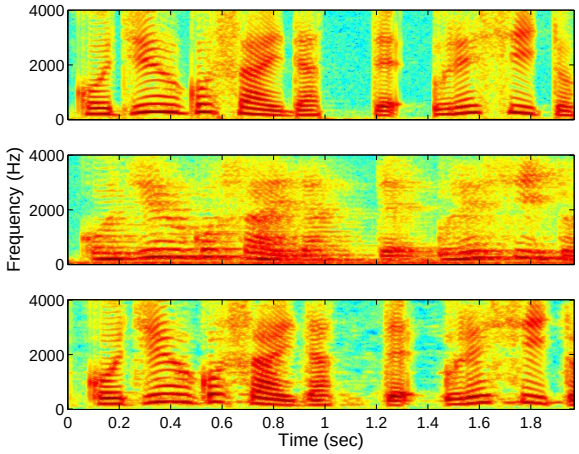

Figure 5: Portions of spectrograms of female clean speech (top), its reverberant version (SD = 5.39 dB, middle), and dereverberated version (SD = 3.26 dB, bottom).

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, "Digital processing of speech signals," Prentice Hall, 1978.

[2] M. K. 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, 1995.

[3] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation of speech signals based on linear prediction," *Proc. ICSLP*, pp. 877–880, 2004.

[4] T. Hikichi, M. Delcroix, and M. Miyoshi, "Blind dereverberation based on estimates of signal transmission channels without precise information on channel order," *Proc. ICASSP*, pp. 1069–1072, 2005.

[5] P. Wheeler, S. Kajita, K. Takeda, and F. Itakura, "Blind deconvolution using information maximization," *Tech. Rep. IEICE*, DSP98–82, SP98–61, pp. 23–29, 1998.

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

[7] 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.

[8] A. Hyvärinen, J. Karhunen, and E. Oja, "Independent component analysis," John Wiley and Sons, 2001.

[9] 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.

[10] ISO/DIS 3382, "Acoustics: measurement of the reverberation time with reference to other acoustical parameters," 1997.