Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction

Tomohiro Nakatani, *Senior Member, IEEE*, Takuya Yoshioka, *Member, IEEE*, Keisuke Kinoshita, *Member, IEEE*, Masato Miyoshi, *Senior Member, IEEE*, and Biing-Hwang Juang, *Fellow, IEEE*

***Abstract*—This paper proposes a statistical model-based speech dereverberation approach that can cancel the late reverberation of a reverberant speech signal captured by distant microphones without prior knowledge of the room impulse responses. With this approach, the generative model of the captured signal is composed of a source process, which is assumed to be a Gaussian process with a time-varying variance, and an observation process modeled by a delayed linear prediction (DLP). The optimization objective for the dereverberation problem is derived to be the sum of the squared prediction errors normalized by the source variances; hence, this approach is referred to as variance-normalized delayed linear prediction (NDLP). Inheriting the characteristic of DLP, NDLP can robustly estimate an inverse system for late reverberation in the presence of noise without greatly distorting a direct speech signal. In addition, owing to the use of variance normalization, NDLP allows us to improve the dereverberation result especially with relatively short (of the order of a few seconds) observations. Furthermore, NDLP can be implemented in a computationally efficient manner in the time–frequency domain. Experimental results demonstrate the effectiveness and efficiency of the proposed approach in comparison with two existing approaches.**

***Index Terms*—Dereverberation, inverse filtering, multichannel linear prediction, speech enhancement, statistical model-based signal processing.**

I. Introduction

A SPEECH signal captured in an enclosed space such as a conference room will inevitably contain reverberant components due to reflections from the walls, floor, or ceiling. These reverberant components are detrimental to the perceived quality of the observed speech signal and often cause serious degradation in many applications such as hands-free teleconferencing and automatic speech recognition (ASR).

The goal of "dereverberation" is to reduce the reverberant components in the acquired signal while preserving the direct signal component prior to its application so as to minimize the aforementioned detrimental effects.

A typical approach to dereverberation is to use microphone array processing [1], [2], which involves estimating the directions of arrival (DOAs) of a direct signal, and enhancing the signal components coming from the source direction by controlling the directivity of the microphone array. For satisfactory dereverberation results, this technique requires a relatively large number of microphones to obtain sufficient directivity gain.

Another approach is based on the suppression of late reverberation in the power spectral domain, for which the reverberation energy is estimated based on the reverberation time of the room [3], [4]. This approach does not involve the phase of the signal, and thus results in relatively robust processing. However, this also limits the effectiveness of the dereverberation [5]. Furthermore, it is not really effective to combine this approach with the other signal processing techniques that try to take advantage of the phase of the signal.

The third approach, which is the main focus of this paper, involves the use of inverse filtering [6], [7]. The aim is to find an inverse filter for the room, which is the cause of the reverberation, and to deconvolve the captured signal with the estimated inverse filter to recover the original signal. When the room transfer functions (RTFs) from the source to the microphones are given, the inverse filter of the RTFs can precisely recover the source signal. The existence of such an inverse filter is guaranteed when the number of microphones exceeds the number of active sources, and the RTFs in individual channels do not share common zeros [6]. In many speech application scenarios, however, we cannot expect the RTFs to be fixed (talkers may change their positions and the room temperature fluctuates continuously) or given in advance. Therefore, we need to estimate the RTFs, or equivalently the inverse filter, from the observed signal. To overcome this problem, there have been many attempts to estimate the inverse filter without prior knowledge of the room impulse response [8]–[14]. There are a number of critically important issues related to the problem, including the following.

1) How to distinguish the inherent characteristics of the speech (e.g., the formant structure of the speech derived from the vocal tract feature) and those of the room acoustics from the reverberant observation.
2) How to regularize the system inversion to make the dereverberation robust against ubiquitous acoustic noise and system modeling errors such as channel order mismatch. These factors are practically omnipresent and in this paper we refer to the deviations in the signal caused by these

Manuscript received November 18, 2009; revised March 13, 2010; accepted May 10, 2010. Date of current version August 13, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Walter L. Kellermann.

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

M. Miyoshi is with the Graduate School of Natural Science and Technology, Kanazawa University, Kakuma, Kanazawa 920-1192 Japan (e-mail: mmiyoshi@t.kanazawa-u.ac.jp).

B. H. Juang is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250 USA (e-mail: juang@ece.gatech.edu).

Digital Object Identifier 10.1109/TASL.2010.2052251

factors as additive noise for the sake of brevity. The strict inverse filter for the given RTFs often amplifies the noise greatly.

3) How to accomplish effective dereverberation with a relatively short (e.g., a few seconds) observation and at a relatively small computational cost.

Existing inverse filtering methods have serious limitations with regard to some of the above issues, and thus dereverberation is still a challenging problem.

In a related problem, a multichannel linear prediction-based channel identification approach has been proposed for identifying wireless communication channels [15], [16], where the source signal is assumed to be an independent and identically distributed (i.i.d.) sequence. This approach is promising in that it can robustly estimate the channels without prior knowledge in the presence of a channel order mismatch and noise. A multistep prediction technique has also been proposed in [17] in a similar context, namely, blind channel identification and equalization. However, unlike wireless communication signals, the speech signal is not an i.i.d. sequence, and thus these methods cannot be directly applied to speech signals. To cope with speech signals, we must develop methods for distinguishing their characteristics from the characteristics of the channels (issue 1 listed in the previous paragraph). This calls for an explicit model of the speech signal.

Following the above observation, we have been employing an approach where the observation process is modeled based on a multichannel linear prediction (i.e., a multichannel autoregressive model) [18]–[20] and the source process is characterized by a statistical source model represented, for example, by a set of spectral prototypes (a vector quantization codebook) [21], [22]. Delayed linear prediction (DLP), which was first proposed as an estimation step for the multistep prediction, was utilized for speech dereverberation in [20]. With DLP, the reverberation is divided into two parts, namely early and late reverberations. It was shown that DLP can suppress the late reverberation effectively without significantly distorting the short time correlations of the speech, with the assumption that speech is stationary. By contrast, following the statistical model-based approach proposed in [22], the use of time-varying speech characteristics with multichannel linear prediction was shown to be useful for effectively reducing not only the late reverberation but also the early reverberation based only on a relatively short (of the order of a few seconds) observation signal. The drawback of this approach is its substantial computing cost, which needs to be reduced for practical applications.

In this paper, we propose a generalization of the DLP-based dereverberation approach, which allows us to handle time-varying signals more appropriately. Originally, linear prediction was proposed for analyzing stationary signals, and this is theoretically inapplicable to time-varying signals. So, the proposed generalization is important in terms of broadening the application areas of linear prediction. This new approach is referred to as variance-normalized delayed linear prediction (NDLP). With NDLP, the time-varying variance of the speech signal is taken into account, and the dereverberation is accomplished based on a maximum-likelihood estimation approach. This approach can also be viewed as a variant of the statistical model based dereverberation approach, where DLP and a simplified statistical speech model are introduced. Thanks to these modifications, NDLP based dereverberation can be easily implemented in the time–frequency domain, which allows us to accomplish statistical model-based dereverberation in a computationally very efficient manner. In summary, the proposed approach can improve the two existing approaches in terms of effectiveness and computational efficiency, respectively.

The significance of variance normalization should not be underestimated. It is not merely a simple scaling procedure; rather, it should be considered a necessary technique for equalizing out the effect of the temporal power level variation inherent in the speech signal during the estimation and the solution of DLP. Without addressing this temporal variation, the DLP would not be considered an effective model and solution for time varying signals.

This paper is an extension of previously presented work [23], where we simply derived the time–frequency domain algorithm of the proposed method and showed its effectiveness in preliminary experiments. This algorithm is often referred to as the weighted prediction error (WPE) method, and has already been integrated with other speech enhancement techniques [24], [25]. In this paper, we first provide a complete formulation of NDLP in the time domain in Section II. Then, in Section III, we provide a thorough mathematical analysis of the behavior and solution of NDLP to show their theoretical justification, which has not been provided until now. In particular, we discuss the relationship between the solution and the statistical characteristics of speech, and the convergence property of NDLP. We show that NDLP may partially distort early reverberation but does not reduce the direct signal component in the dereverberated signal. In Section IV, we present the frequency domain implementation of the proposed method, which is identical to WPE. With this implementation, dereverberation can be achieved very efficiently based on a source model that can precisely represent the characteristics of any time-varying power spectrum. In Section V, we describe comparison experiments that confirm the effect of the variance normalization, computational efficiency, and noise robustness of NDLP. Our concluding remarks are presented in Section VI.

II. Definition of Variance-Normalized Delayed Linear Prediction

Suppose a single speech source is captured by L_m microphones ($L_m > 1$) with a certain level of acoustic noise. Let s_t, $x_t^{(m)}$, and $b_t^{(m)}$ be digitized sequences of the source, observed, and noise signals, respectively, where t and m are the sample and microphone (channel) indices, respectively. The room impulse response (RIR) of length L_h from the source to the mth microphone is denoted by $\overline{h}^{(m)} = [h_0^{(m)}, h_1^{(m)}, \ldots, h_{L_h-1}^{(m)}]^T$ where T indicates a matrix transposition operation. Then, the observed signal can be modeled by

$$x_t^{(m)} = \sum_{k=0}^{L_h-1} h_k^{(m)} s_{t-k} + b_t^{(m)}. \tag{1}$$

In this paper, we define the goal of dereverberation as to estimate a dereverberation filter that produces an enhanced signal

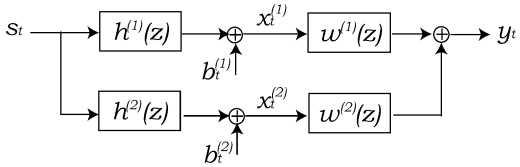

Fig. 1. Observation and dereverberation models. $h^{(m)}(z)$ and $w^{(m)}(z)$ are the room transfer function (RTF) and the transfer function of the dereverberation filter.

y_t that contains less reverberation than the observed signal $x_t^{(m)}$ without greatly increasing the acoustic noise level. The enhanced signal y_t is obtained as

$$y_t = \sum_{m=1}^{L_m} \sum_{k=0}^{L_w-1} w_k^{(m)} x_{t-k}^{(m)} \tag{2}$$

where $\bar{w}_t^{(m)} = [w_0^{(m)}, w_1^{(m)}, \ldots, w_{L_w-1}^{(m)}]^T$ is a dereverberation filter of length L_w. Fig. 1 shows the observation and dereverberation models represented by (1) and (2).

In the following, we assume that there are two microphones, namely $L_m = 2$, for the sake of simplicity.[1] In addition, we focus on noise-free observations, where $b_t^{(m)} = 0$ in (1), in this section, and we extend the discussion to cope with noisy observations, namely $b_t^{(m)} \neq 0$, in Sections III-C and IV-D.

A. Observation Model With Delayed Linear Prediction

The reverberant observed signal in (1) without noise can be decomposed into three parts, a direct signal, and early and late reverberation parts. In this paper, the sum of the former two parts is taken as the signal to be obtained, which is referred to as the desired signal and denoted by $d_t^{(m)}$, and the late reverberation part is taken as the signal to be eliminated, denoted by $r_t^{(m)}$. Their relationship can be expressed as

$$x_t^{(m)} = d_t^{(m)} + r_t^{(m)} \tag{3}$$

where

$$d_t^{(m)} = \sum_{k=0}^{D-1} h_k^{(m)} s_{t-k} \quad \text{and } r_t^{(m)} = \sum_{k=D}^{L_h-1} h_k^{(m)} s_{t-k} \tag{4}$$

where D is the sample index by which we separate the room impulse response into early and late reverberation parts, and which we hereafter refer to as the prediction delay. The meaning of "prediction delay" is provided in the next paragraph.

When we assume that the RTFs in different channels, defined corresponding to $\bar{h}^{(m)}$, do not share common zeros, the relationship between the speech signal and the observed signals in (1) can be rewritten (as shown in [17] and also in Appendix A) as

$$x_t^{(m)} = \sum_{m'=1}^{L_m} \sum_{k=0}^{L_c-1} x_{t-D-k}^{(m')} c_k^{(m',m)} + d_t^{(m)}. \tag{5}$$

[1]It is straightforward to extend the discussion in this paper to the use of more than two microphones. In addition, the proposed method is shown to be effective based on our experience even with a single microphone. However, it should be noted that there may be no exact inverse filter for the single-channel case.

We also use vector representations of the above observation model to simplify the following discussion:

$$x_t^{(m)} = (\bar{c}^{(m)})^T \bar{x}_{t-D} + d_t^{(m)} \tag{6}$$

where

$$\begin{aligned}
\bar{x}_t &= \left[(\bar{x}_t^{(1)})^T, (\bar{x}_t^{(2)})^T\right]^T \quad (2L_c \times 1) \\
\bar{x}_t^{(m)} &= \left[x_t^{(m)}, x_{t-1}^{(m)}, \ldots, x_{t-L_c+1}^{(m)}\right]^T \quad (L_c \times 1) \\
\bar{c}^{(m)} &= \left[(\bar{c}^{(1,m)})^T, (\bar{c}^{(2,m)})^T\right]^T \quad (2L_c \times 1) \\
\bar{c}^{(m',m)} &= \left[c_1^{(m',m)}, c_2^{(m',m)}, \ldots, c_{L_c}^{(m',m)}\right]^T, \quad (L_c \times 1).
\end{aligned}$$

In (5) and (6), the current observed signal $x_t^{(m)}$ is predicted by convolving a series of past observed signals $\bar{x}_{t-D}$ with the regression coefficient $c^{(m)}$ of the order of L_c and the prediction residual is the desired signal $d_t^{(m)}$.

With estimated regression coefficients $\hat{\bar{c}}^{(m)}$ and according to (6), the desired signal can be estimated as

$$\hat{d}_t^{(m)} = x_t^{(m)} - \left(\hat{\bar{c}}^{(m)}\right)^T \bar{x}_{t-D}. \tag{7}$$

Therefore, the dereverberation can be accomplished by obtaining an appropriate regression coefficient $\hat{\bar{c}}^{(m)}$ from the observed signal. Hereafter, as in (6), a scalar variable is represented by a lowercase symbol, and a vector variable is represented by a lowercase symbol with a bar "$\bar{\cdot}$." In addition, as in (7), we attach a hat "$\hat{}$" to a symbol denoting an estimated variable.

In this paper, we use (6) as the observation model for DLP, where the latest D samples of the past observed signal are not used for the prediction of the current observed signal. Furthermore, in the probabilistic formulation described in the following, we interpret (6) as follows.

- Provided $\bar{x}_{t-D}$ and $\bar{c}^{(m)}$ are given, the probabilistic uncertainty of the current observed signal $x_t^{(m)}$ is derived only from that of $d_t^{(m)}$, and $x_t^{(m)}$ does not depend on the latest D samples of the observed signal $x_{t'}^{(m)}$ for $t > t' > t - D$.

B. Definition of Likelihood Function

With the proposed approach, a log likelihood function based on a generative model composed of the speech and the observation models is introduced as the optimization criterion to determine the regression coefficient $\bar{c}^{(m)}$. By setting the parameter set at $\theta = \{\bar{c}^{(1)}, \bar{c}^{(2)}\}$, the log likelihood function is defined as follows:

$$\mathcal{L}(\theta) = \log p(\bar{x}_{\langle \mathcal{T}, -\infty \rangle}; \theta). \tag{8}$$

Here, $p(\bar{x})$ is the probability density function (pdf) of $\bar{x}$, and $\bar{x}_{\langle t_2, t_1 \rangle}$ is a vector containing a time series of the observed signals of all the channels from sample indices t_1 to t_2 defined as

$$\begin{aligned}
\bar{x}_{\langle t_2, t_1 \rangle} &= \left[\left(\bar{x}_{\langle t_2, t_1 \rangle}^{(1)}\right)^T, \left(\bar{x}_{\langle t_2, t_1 \rangle}^{(2)}\right)^T\right]^T \\
\bar{x}_{\langle t_2, t_1 \rangle}^{(m)} &= \left[x_{t_2}^{(m)}, x_{t_2-1}^{(m)}, \ldots, x_{t_1}^{(m)}\right]^T
\end{aligned}$$

$\mathcal{T}$ is the largest sample index of the observed signal. As shown in Appendix B, the log likelihood function can be approximately rewritten as

$$\mathcal{L}(\theta) \approx \sum_{t=1}^{\mathcal{T}} \log p\left(x_t^{(1)} \middle| \bar{x}_{t-D}; \theta\right) \tag{9}$$

$$= \sum_{t=1}^{\mathcal{T}} \log p\left(d_t^{(1)} = x_t^{(1)} - \left(\bar{c}^{(1)}\right)^T \bar{x}_{t-D}; \theta\right). \tag{10}$$

Here, $p(d_t^{(1)})$ is the marginal pdf of the desired signal $d_t^{(1)}$ at sample index t, where $d_{t'}^{(1)}$ for all $t' \neq t$ are marginalized. The above log likelihood function can be specified simply by giving the definition of the pdf $p(d_t^{(1)})$. It should be noted that instead of using a marginal pdf for each speech sample as the pdf for the desired signal in (10), we may use a multivariate pdf for each short time frame as discussed in [22], which allows us to handle short time correlations of speech more appropriately. Nevertheless, in this paper, we adopt the simpler model, namely the marginal pdf, to define NDLP and to investigate its characteristics in detail. As shown later by mathematical analysis, even with this simple model, NDLP can approximately handle time-varying speech characteristics. Furthermore, with the time–frequency domain implementation discussed in Section IV, even this simple model can precisely represent the time-varying characteristics of speech, including the spectral shapes and powers, and enable NDLP to work more effectively for speech dereverberation than with the time domain implementation.

According to (10) (and its variant with a multivariate source pdf), the dereverberation can be viewed as a problem of finding regression coefficients that make the resultant desired signal most likely to be the desired signal in terms of the pdf $p(d_t^{(1)})$ in (10). It may also be noteworthy that (10) only contains $\bar{c}^{(1)}$ and does not contain $\bar{c}^{(2)}$. This results from the noise-free assumption that we introduced in this section, with which $\bar{c}^{(1)}$ is completely determined independently of $\bar{c}^{(2)}$. So, in the following, we disregard the optimization of $\bar{c}^{(2)}$ without loss of generality.[2]

C. Speech Signal Model

In this section, we introduce several assumptions regarding the characteristics of the speech signal. First, the speech signal is assumed to be a quasi-stationary process, having a correlation only within a short time frame of the order of tens of milliseconds, and its correlation property may vary over different short time frames. The characteristics of speech are assumed as follows.

1) Short-time Gaussianity: within each short time frame of length L_f, the speech signal s_t is a realization of a stationary univariate Gaussian process with a mean zero and a covariance matrix $R_t = E\{\bar{\bar{s}}_t^T \bar{\bar{s}}_t\}$ where $\bar{\bar{s}}_t^{(1)} = [s_{t-L_f/2+1}, \ldots, s_{t+L_f/2}]^T$. Then, the desired signal $d_t^{(1)}$ is also a Gaussian process based on (4) because the weighted sum of random variables following a Gaussian process also follows a Gaussian process. So, the marginal pdf of $d_t^{(1)}$ can be expressed as

$$p\left(d_t^{(1)}\right) = \mathcal{N}\left(d_t^{(1)}; 0, \sigma_t^2\right) \tag{11}$$

where $\sigma_t^2 = E\{|d_t^{(1)}|^2\}$.

2) Local mixing property: two speech samples at sample indices t and t' are mutually uncorrelated when $|t - t'| > \delta$ for a certain constant $\delta > 0$, which is represented as

$$E\{s_t s_{t'}\} = 0 \quad \text{for } |t - t'| > \delta. \tag{12}$$

3) Time-varying variance: σ_t^2 is constant within a short time frame, and varies over different frames. It can be well estimated by the time average of the desired signal within a short time frame as

$$\sigma_t^2 \approx \frac{1}{L_f} \sum_{t'=t-L_f/2+1}^{t+L_f/2} |d_{t'}^{(1)}|^2. \tag{13}$$

Because our focus in this paper is mainly on a representation of the time-varying characteristics of speech, in the above model we adopted a rather simple model of the distribution for short time frames, namely a Gaussian distribution. Discussions of more appropriate model distributions of the frames will constitute future work.

D. Solution Based on Likelihood Maximization

Let $\bar{\sigma}^2 = \{\sigma_1^2, \sigma_2^2, \ldots\}$ be a set of σ_t^2 for all sample indices t, and $\theta' = \{\bar{c}^{(1)}, \bar{\sigma}^2\}$ be the parameter set to be estimated. Then, the log likelihood function (10) can be rewritten, according to (11), as

$$\mathcal{L}(\theta') = \sum_{t=1}^{\mathcal{T}} \log \mathcal{N}\left(d_t^{(1)} = x_t^{(1)} - (\bar{c}^{(1)})^T \bar{x}_{t-D}; 0, \sigma_t^2\right)$$

$$= -\frac{1}{2} \sum_{t=1}^{\mathcal{T}} \frac{\left|x_t^{(1)} - \left(\bar{c}^{(1)}\right)^T \bar{x}_{t-D}\right|^2}{\sigma_t^2} - \frac{1}{2} \sum_{t=1}^{\mathcal{T}} \log \sigma_t^2 + \text{const.} \tag{14}$$

The dereverberation with likelihood maximization is accomplished by finding θ' that maximizes the above equation. In the following, we set a certain lower bound $\epsilon > 0$ for $\hat{\sigma}_t^2$ as $\hat{\sigma}_t^2 \geq \epsilon$ to avoid zero division.

Although it is difficult to obtain a closed form solution for the maximization of (14), we can iteratively increase the objective by alternately updating $\hat{\bar{c}}^{(1)}$ and $\hat{\bar{\sigma}}^2$. The resultant optimization algorithm can be summarized as follows (see also Fig. 2)[3].

1) Initialize $\hat{\sigma}_t^2$ as

$$\hat{\sigma}_t^2 = \max\left\{\frac{1}{L_f} \sum_{t'=t-L_f/2+1}^{t+L_f/2} \left|x_t^{(1)}\right|^2, \epsilon\right\}. \tag{15}$$

[2]If we estimate $\bar{c}^{(m)}$ for all m values independently and apply it to the observed signal, we obtain a multichannel dereverberated signal. So, further improvement of the resultant speech quality may be obtained with multichannel speech enhancement postprocessing techniques, such as delay and sum beamformer and source separation [20], [24].

[3]It is also effective to apply pre-whitening to the observed signal before estimating the regression coefficients in order to reduce the effect of correlation included in speech in the same way as in [20]. We adopt this preprocessing for testing time domain algorithms in our experiments.

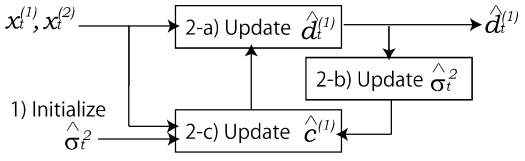

Fig. 2. Processing flow of the variance-normalized delayed linear prediction in the time domain.

2) Repeat the following until convergence.
   a) Update $\hat{\bar{c}}^{(1)}$ as follows:

$$\hat{\bar{c}}^{(1)} = \hat{\Phi}^{+}\hat{\phi} \tag{16}$$

where "+" is the Moore–Penrose pseudo-inverse and

$$\hat{\Phi} = \sum_{t=1}^{\mathcal{T}} \frac{\bar{x}_{t-D}\bar{x}_{t-D}^{T}}{\hat{\sigma}_t^2} \tag{17}$$

$$\hat{\phi} = \sum_{t=1}^{\mathcal{T}} \frac{\bar{x}_{t-D}x_t^{(1)}}{\hat{\sigma}_t^2}. \tag{18}$$

   b) Update $\hat{d}_t^{(1)}$ as $\hat{d}_t^{(1)} = x_t^{(1)} - (\hat{\bar{c}}^{(1)})^{T}\bar{x}_{t-D}$.
   c) Update $\hat{\sigma}_t^2$ as follows:

$$\hat{\sigma}_t^2 = \max\left\{\frac{1}{L_f}\sum_{t'=t-L_f/2+1}^{t+L_f/2}\left|\hat{d}_t^{(1)}\right|^2, \epsilon\right\}. \tag{19}$$

It is guaranteed that the log likelihood given by (14) increases monotonically with the above update procedures in step 2) because (16) and (19) are derived as transformations that maximize (14) with fixed $\hat{\sigma}_t^2$ and $\hat{\bar{c}}^{(1)}$, respectively. The convergence of the above iterative optimization is also guaranteed because (14) is upper-bounded as long as $\hat{\sigma}_t^2$ is updated by (19) with the lower bound ϵ. As a result of the iteration, $\hat{d}_t^{(1)}$ is the estimated desired signal, or the dereverberated speech signal.

It is important to note that the update equation for $\bar{c}^{(1)}$ in (16) can be viewed as a variant of a least square solution to a linear prediction analysis, commonly referred to as the covariance method. A unique characteristic of (16) is that the covariance matrix $\hat{\Phi}$ and the covariance vector $\hat{\phi}$ of the observed signal are calculated while being normalized by the time-varying variance of the desired signal $\hat{\sigma}_t^2$. The other characteristic is that the prediction delay "D" is used in the calculation of the covariance vector $\hat{\phi}$. Because of these characteristics, we refer to this dereverberation approach as variance-normalized delayed linear prediction. One may speculate that the proposed method could also be implemented using the correlation method by approximating (17) employing a sum of the variance-normalized correlation matrices, thereby allowing us to use efficient optimization algorithms such as the Levinson–Durbin recursion to obtain the solution. However, this results in substantial degradation in the computational accuracy because σ_t^2 changes quickly and thus the covariance matrix cannot be approximated by a correlation matrix. So, we do not employ the correlation method in this paper.

In the rest of this paper, we fix the number of iterations in the above optimization algorithm at one; that is, we simply substitute the variance of the observed signal for that of the desired signal. In our preliminary experiments it turns out that this substitution works very well while the iterative optimization does not always improve the dereverberation accuracy. This is probably because the assumption regarding the time-varying nature of speech is oversimplified, that is, the variance is assumed to be time varying with no further assumptions as regards its time envelope characteristics and the spectral shape is assumed to be flat and not to change. In contrast, it has already been shown that the same type of iterative optimization improves the dereverberation accuracy when we adopt more precise source models [22]. Therefore, we presume that the appropriateness of the estimation of σ_t^2 should be discussed in relation to the more precise time-varying characteristics of the variance. This is beyond the scope of this paper.

As defined in the above optimization algorithm, this paper focuses on batch processing, where all the observed signals are assumed to be given in advance for the optimization. With regard to the extension of this approach to cope with adaptive processing, which is important for real-time applications such as teleconferencing, preliminary discussions can be found, for example, in [26], [27].

III. Analysis of Solution With Variance-Normalized Delayed Linear Prediction

In this section, we analyze in detail the characteristics of the solution obtained based on NDLP. We show the following characteristics regarding the solution.

- The solution depends only on the statistical characteristics of speech, the early reverberation, and the estimated variance $\hat{\sigma}_t^2$. The late reverberation included in the observed signal is completely eliminated from the solution. (see Section III-A).
- With sufficiently long observations, the estimated desired signal results in the sum of the direct signal and the early reverberation where only part of the early reverberation is distorted due to the statistical characteristics of the speech. (see Section III-B).
- The variance normalization is expected to improve the accuracy of the dereverberation when the estimated variance of the desired signal $\hat{\sigma}_t^2$ appropriately represents the speech characteristics. (see Section III-B2).
- The robustness of the algorithm in the presence of noise is also discussed. (see Section III-C)

Hereafter, we assume that an estimate of the variance of the desired signal $\hat{\sigma}_t^2$ is obtained from the observed signal as in (15), and is fixed during the maximization of the log likelihood function $\mathcal{L}(\theta')$ in (14).

A. Relationship Between True and Estimated Desired Signals

Based on a property of the Moore–Penrose pseudo-inverse, (16) gives the regression coefficient vector $\hat{\bar{c}}^{(1)}$ that has the minimum norm $|\hat{\bar{c}}^{(1)}|^2$ of those that maximize (14) when the values of $\hat{\sigma}_t^2$ for all t are fixed. Hereafter, this solution is referred to as the minimum-norm solution. We start this section by deriving a different form of this solution. We use true values of the room impulse responses and assume that $\bar{s}_t$ and $\bar{d}_t$ represent the true

values of the unknown speech and desired signals, respectively. First, (1) can be rewritten in a matrix form as

$$\bar{x}_t = H\bar{s}_t \tag{20}$$

where

$$\begin{aligned}\bar{s}_t &= [s_t, \ldots, s_{t-L+1}]^T, \quad (L \times 1)\\ H &= \left[(H^{(1)})^T, (H^{(2)})^T\right]^T, \quad (2L_c \times L)\end{aligned} \tag{21}$$

$L = L_c + L_h - 1$, and $H^{(m)}$ is a convolution matrix of size $L_c \times L$ defined corresponding to the room impulse response $\bar{h}^{(m)} = [h_0^{(m)}, \ldots, h_{L_h-1}^{(m)}]^T$. Note that H is the full column rank when the room transfer functions do not share common zeros [6]. Next, let the following be vectors of length L containing the earlier and later parts of the room impulse response, $\hat{h}_e^{(1)}$ and $\hat{h}_l^{(1)}$, respectively, and referred to as early and late room impulse responses hereafter.

$$\bar{h}_e^{(1)} = \left[h_0^{(1)}, h_1^{(1)}, \ldots, h_{D-1}^{(1)}, 0, \ldots, 0\right]^T \tag{22}$$

and

$$\bar{h}_l^{(1)} = \left[h_D^{(1)}, h_{D+1}^{(1)}, \ldots, h_{L_h-1}^{(1)}, 0, \ldots, 0\right]^T. \tag{23}$$

With this notation, (4) can be rewritten as

$$d_t^{(1)} = \left(\bar{h}_e^{(1)}\right)^T \bar{s}_t, \quad \text{and } r_t^{(1)} = \left(\bar{h}_l^{(1)}\right)^T \bar{s}_{t-D}. \tag{24}$$

Using (3), (20), and (24), the first term of the log likelihood function (14) can be rewritten as

$$\begin{aligned}&\mathcal{L}_1(\theta')\\ &= -\frac{1}{2}\sum_{t=1}^{\mathcal{T}} \frac{\left|x_t^{(1)} - \left(\bar{c}^{(1)}\right)^T \bar{x}_{t-D}\right|^2}{\hat{\sigma}_t^2}\\ &= -\frac{1}{2}\sum_{t=1}^{\mathcal{T}} \frac{\left|d_t^{(1)} + \left(\bar{h}_l^{(1)}\right)^T \bar{s}_{t-D} - \left(\bar{c}^{(1)}\right)^T H\bar{s}_{t-D}\right|^2}{\hat{\sigma}_t^2}.\end{aligned} \tag{25}$$

Then, the minimum-norm solution that maximizes (25) can be derived, assuming $\sum_{t=1}^{\mathcal{T}} \bar{s}_{t-D}\bar{s}_{t-D}^T/\hat{\sigma}_t^2$ to be non-singular, as (see Appendix C)

$$\hat{\bar{c}}^{(1)} = H(H^TH)^{-1}\left(\bar{h}_l^{(1)} + \bar{\xi}\right) \tag{26}$$

where

$$\bar{\xi} = \left(\sum_{t=1}^{\mathcal{T}} \frac{\bar{s}_{t-D}\bar{s}_{t-D}^T}{\hat{\sigma}_t^2}\right)^{-1} \sum_{t=1}^{\mathcal{T}} \frac{\bar{s}_{t-D}d_t^{(1)}}{\hat{\sigma}_t^2}. \tag{27}$$

Substituting (26) into (7), we can rewrite the estimated desired signal as

$$\hat{d}_t^{(1)} = d_t^{(1)} - \bar{\xi}^T\bar{s}_{t-D} \tag{28}$$

where it is represented in terms of its true value and the estimation error. The second term of the above equation is the estimation error, and it is examined carefully for an analysis of the solution in the following. By substituting (4) into (27), $\bar{\xi}$ can also be represented by a linear combination of the early room impulse response as

$$\begin{aligned}\bar{\xi} &= \sum_{k=0}^{D-1} \bar{\xi}_k h_k^{(1)}\\ \bar{\xi}_k &= \left(\sum_{t=1}^{\mathcal{T}} \frac{\bar{s}_{t-D}\bar{s}_{t-D}^T}{\hat{\sigma}_t^2}\right)^{-1} \sum_{t=1}^{\mathcal{T}} \frac{\bar{s}_{t-D}s_{t-k}}{\hat{\sigma}_t^2}\end{aligned} \tag{29}$$

and $\bar{\xi}_k$ in (29) corresponds to the value obtained from a maximization operation defined as

$$\bar{\xi}_k = \arg\max_{\bar{\xi}_k'} \left(-\frac{1}{2}\sum_{t=1}^{\mathcal{T}} \frac{\left|s_{t-k} - \bar{\xi}_k'^T\bar{s}_{t-D}\right|^2}{\hat{\sigma}_t^2}\right). \tag{30}$$

This means that $\bar{\xi}_k$ corresponds to the regression coefficients that best predict the past speech sample at $t-k$ from the past speech samples before sample index $t-D$ based on the minimum prediction error criterion with variance normalization. Substituting (29) into (28), we obtain

$$\hat{d}_t^{(1)} = \sum_{k=0}^{D-1} h_k^{(1)}(s_{t-k} - \tilde{s}_{t,k}) \tag{31}$$

$$\tilde{s}_{t,k} = \bar{\xi}_k^T\bar{s}_{t-D} \tag{32}$$

where $\tilde{s}_{t,k}$ is the predicted clean speech signal at sample index $t-k$ using $\bar{s}_{t-D}$ and $\bar{\xi}_k$ obtained by (30). From (31), it is shown that the estimated desired signal $\hat{d}_t^{(1)}$ is equal to the convolution of the early room impulse response and the difference between the true and predicted clean speech signals, s_{t-k} and $\tilde{s}_{t,k}$, respectively. Note that all the errors in the estimated desired signal are caused by the predicted clean speech signal $\tilde{s}_{t,k}$, and $\tilde{s}_{t,k}$ would be zero in the sense of expectation if the speech signal were uncorrelated and zero-mean. Finally, we can conclude the following from the above equation.

- All the late reverberation included in the observed signal is eliminated from the estimated desired signal.
- The estimation error occurs from the predicted clean speech signal $\tilde{s}_{t,k}$, namely a certain part of the clean speech signal s_{t-k} that can be predicted by the past clean speech signal, $\bar{s}_{t-D}$. Interestingly, $\tilde{s}_{t,k}$ depends only on the statistical characteristics of the clean speech signal, and not on the late room impulse responses.
- Because the order of the regression coefficient is much larger than the length of a short time frame, the estimation error may contain the weighted sum of the clean speech signals over a long duration.

Note that on the assumption that the signal is stationary we can derive the same kind of DLP characteristics as discussed above. In other words, NDLP can be viewed as an extension of DLP that is suitable for analyzing the signals with time-varying source variances in the above sense.

B. Error Analysis Based on Speech Characteristics

In this subsection, we further analyze the property of the estimation error, $\tilde{s}_{t,k}$, which is caused by the correlation in-

herent in speech, based on the speech characteristics defined in Section II-C.

1) Convergence Behavior: Here, we investigate the convergence behavior of NDLP when the observation is sufficiently long. For this investigation, in relation to assumption 2 introduced in Section II-C, we introduce the following assumption, that is, the time average shown below on the left-hand side converges to zero as the observation becomes longer, (i.e., $\mathcal{T} \to \infty$)

$$\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \frac{s_{t-k} s_{t-k'}}{\hat{\sigma}_t^2} \to 0 \quad \text{for } |k' - k| > \delta \tag{33}$$

where k and k' are any constant integers. The appropriateness of this assumption can be roughly explained as follows: let us first discretize the values of $\hat{\sigma}_t^2$ into a set of $N_g (\ll \mathcal{T})$ positive values $\{g_1, g_2, \ldots, g_{N_g}\}$ and let $\Omega(g_n)$ be a set of sample indices t for which $\hat{\sigma}_t^2$ is categorized into g_n, and $L_{\Omega(g_n)}$ be the number of elements in $\Omega(g_n)$. Then, the above time average can be approximated as follows:

$$\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \frac{s_{t-k} s_{t-k'}}{\hat{\sigma}_t^2} \approx \sum_{n=1}^{N_g} \frac{L_{\Omega(g_n)}}{\mathcal{T}} G_n$$

$$G_n = \frac{1}{L_{\Omega(g_n)}} \sum_{t \in \Omega(g_n)} \frac{s_{t-k} s_{t-k'}}{g_n}$$

G_n in the above equation presumably approaches zero as the observation lengthens because s_{t-k} and $s_{t-k'}$ are uncorrelated for $|k - k'| > \delta$, and thus (33) is also expected to converge to zero.

Then, the summation term within the parentheses in (30) can be rewritten as

$$\sum_{t=1}^{\mathcal{T}} \frac{|s_{t-k} - \bar{\xi}_k'^T \bar{s}_{t-D}|^2}{\hat{\sigma}_t^2}$$

$$= \sum_{t=1}^{\mathcal{T}} \frac{|s_{t-k}|^2}{\hat{\sigma}_t^2} - 2\bar{\xi}_k^T \sum_{t=1}^{\mathcal{T}} \frac{s_{t-k} \bar{s}_{t-D}}{\hat{\sigma}_t^2} + \bar{\xi}_k^T \sum_{t=1}^{\mathcal{T}} \frac{\bar{s}_{t-D} \bar{s}_{t-D}^T}{\hat{\sigma}_t^2} \bar{\xi}_k.$$

Dividing both sides in the above equation by the total number of samples $\mathcal{T}$, the second term on the right-hand side converges to zero for $|D - k| > \delta$ according to assumption (33). In addition, $\sum_{t=1}^{\mathcal{T}} \bar{s}_{t-D} \bar{s}_{t-D}^T / \hat{\sigma}_t^2$ in the third term is assumed to be positive definite. Therefore, $\bar{\xi}_k$ that maximizes (30) is expected to converge to zero. As a consequence, we can derive the following when $\mathcal{T} \to \infty$:

$$\bar{\xi}_k \to 0 \quad \text{for } |D - k| > \delta. \tag{34}$$

With this result, when the observation is sufficiently long, (31) can be rewritten as

$$\hat{d}_t^{(1)} = \sum_{k=0}^{D-\delta} h_k^{(1)} s_{t-k} + \sum_{k=D-\delta+1}^{D} h_k^{(1)} (s_{t-k} - \tilde{s}_{t,k}). \tag{35}$$

From this equation, we can conclude the following.

- Under assumption (33), in the desired signal estimated based on a sufficiently long observation, the direct signal remains unchanged when the prediction delay D is larger than the duration δ, over which neither of the two speech samples have correlations. Only part of the early reverberation is distorted owing to the speech signal's inherent correlation.

This guarantees that at least the direct signal remains in the estimated desired signal. The estimation error is only derived from the distortion of the early reverberation.

2) Effect of Variance Normalization: As will be shown by the experiments reported in this paper, an advantage of using variance normalization for speech dereverberation is that we can improve the accuracy of the dereverberation, especially when the observation is short. We consider that this advantage results from the accuracy of the speech model. That is, the dereverberation with variance normalization is accomplished by finding the regression coefficients that transform the observed signal to one that is more likely to be the desired signal in terms of the model pdf. Therefore, when we have good estimates for σ_t^2, we can generally obtain better dereverberation results than without the use of variance normalization where we simply assume σ_t^2 to be constant [20].

For example, this advantage can be intuitively confirmed by comparison with a case without the variance normalization as follows.

- Without the variance normalization, where we set $\hat{\sigma}_t^2 = 1$ for all t, the log likelihood function (10) can be rewritten as

$$\mathcal{L}(\theta') = -\frac{1}{2} \sum_{t=1}^{\mathcal{T}} |\hat{d}_t|^2. \tag{36}$$

  On the other hand, the variance σ_t^2 of the speech signal varies greatly with t. So, by simply maximizing the above function, short time frames with larger σ_t^2 values have a larger impact on the likelihood maximization compared with those with smaller σ_t^2 values. As a result, the estimated coefficients tend to be effective in reducing the energy of the frames with higher σ_t^2 values while they may not be effective in reducing the energy of the frames with lower σ_t^2 values. They may even increase the reverberation energy for frames with smaller σ_t^2 values. This is particularly undesirable for dereverberation because reverberation is generally perceived in frames with relatively small σ_t^2 values.
- With variance normalization, the log likelihood function can be rewritten as

$$\mathcal{L}(\theta') = -\frac{1}{2} \sum_{t=1}^{\mathcal{T}} \frac{|\hat{d}_t|^2}{\hat{\sigma}_t^2} - \frac{N}{2} \sum_{t=1}^{\mathcal{T}} \log \hat{\sigma}_t^2. \tag{37}$$

  Here, the squared prediction error of each sample is normalized by the source variance $\hat{\sigma}_t^2$, and thus the scale of the first term on the right-hand side is almost constant over different short time frames. This can equalize the impacts of frames with different σ_t^2 values on the log likelihood function. As a result, the log likelihood function consists of equal contributions from the individual frames, and thus it is less likely that the energy of the signals in frames with smaller σ_t^2 values will be increased by the dereverberation.

C. Robustness in the Presence of Noise

Here, we briefly discuss the fact that the proposed method works robustly even in the presence of moderate noise without any modification of the algorithm derived in Section II. When the noise signal in (1) is taken into account, the observed signal can be decomposed into noise-free reverberant signal z_t and noise b_t as $x_t = z_t + b_t$, where we assume z_t and b_t are mutually uncorrelated, namely $E\{z_t b_{t'}\} = 0$ for any (t, t'). Then, the log likelihood function (14) can be divided into three parts as

$$\mathcal{L}(\theta') = -\frac{1}{2}\sum_{t=1}^{\mathcal{T}} \frac{\left| z_t^{(1)} - \left(\bar{c}^{(1)}\right)^T \bar{z}_{t-D} \right|^2}{\hat{\sigma}_t^2} - \frac{1}{2}\sum_{t=1}^{\mathcal{T}} \frac{\left| b_t^{(1)} - \left(\bar{c}^{(1)}\right)^T \bar{b}_{t-D} \right|^2}{\hat{\sigma}_t^2} - \frac{1}{2}\sum_{t=1}^{\mathcal{T}} \log \hat{\sigma}_t^2. \quad (38)$$

The dereverberation is achieved by maximizing the first and third terms simultaneously as discussed above, while the resultant noise level after dereverberation is reduced by maximizing the second term because the numerator is the power of the resultant noise. The likelihood function is maximized by jointly maximizing all the terms. In this case, the dereverberation effect may deteriorate due to the second term. However, when the noise level is not very high, we can presume that the first and third terms can be reduced to some extent, and thus the dereverberation is still effective. In addition, the noise level is not greatly increased as a result of the dereverberation owing to the inclusion of the second term in the log likelihood function.

It is noteworthy that the introduction of the prediction delay "D" is expected to work favorably for dereverberation with noise. This is because the level of the desired signal, namely the numerator of the first term in the log likelihood function (38), can be set fairly high relative to that of the noise, namely the second term.

IV. Time–Frequency Domain Implementation

In this section, we discuss a variant of NDLP, which is implemented in the time–frequency domain. With this implementation, the dereverberation processing is decomposed into individual frequency bins and is performed by NDLP independently at each frequency (see Fig. 3). We expect to obtain the following advantages by this modification.

1) The computational cost can be greatly reduced. This is because the size of the variance-normalized covariance matrices and vectors, (17) and (18), can be greatly reduced thanks to the frequency decomposition.
2) Any time-varying power spectrum can be represented precisely by the simple source model used by NDLP because the speech is modeled independently in each frequency bin.
3) It is relatively easy to integrate NDLP in a unified manner with the other commonly used speech enhancement techniques that operate in the time–frequency domain, such as blind source separation and Wiener filtering.

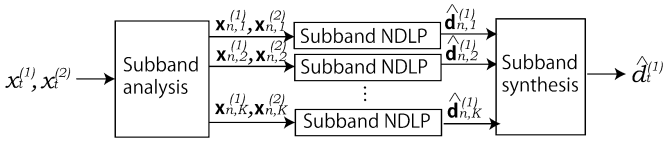

Fig. 3. Processing flow of the variance-normalized delayed linear prediction in the time–frequency domain. Each "Subband NDLP" block executes dereverberation based on NDLP in each subband.

A. Approximation of Time Domain Convolution in Time–Frequency Domain

One simple way of transforming a time domain algorithm into a time–frequency domain algorithm is to use subband decomposition. For this transformation, we adopt the complex subband processing proposed in [28]. Short-time Fourier transformation (STFT) can also be used for this purpose [23].

Using the subband decomposition (or the STFT), the time-domain observation model (1) can be represented in the time–frequency domain as separate observation models in individual subbands. It is defined as

$$\mathbf{x}_{n,f}^{(m)} = \sum_{k=0}^{L_{h'}-1} \left(\mathbf{h}_{k,f}^{(m)}\right)^* \mathbf{s}_{n-k,f} + \mathbf{e}_{n,f}^{(m)} + \mathbf{b}_{n,f}^{(m)} \quad (39)$$

where $\mathbf{x}_{n,f}^{(m)}$, $\mathbf{s}_{n,f}^{(m)}$, and $\mathbf{b}_{n,f}^{(m)}$ are the time–frequency domain signals corresponding to the time domain signals, x_t, s_t, and b_t, respectively, and $\mathbf{h}_{n,f}$ is the time–frequency domain approximation of the room impulse response defined based on the subband processing algorithm. Hereafter, we use boldface symbols to represent variables in the time–frequency domain. n, f, and m are frame, frequency, and channel indices, and $L_{h'}$ is the length of the room impulse response in each subband. When the over-sampling rate of the subband decomposition is F and the number of subbands is K, $L_{h'} = \lceil FL_h/K \rceil$, where $\lceil x \rceil$ denotes the minimum integer that is larger than x. "$*$" denotes a complex conjugate operation. Because the time–frequency domain approximation of the convolution inevitably contains a certain amount of modeling error, it is represented as $\mathbf{e}_{n,f}^{(m)}$ in the above model.

B. Observation Model With Delayed Linear Prediction

By first disregarding the noise and the modeling error terms, $\mathbf{e}_{n,f}^{(m)}$ and $\mathbf{b}_{n,f}^{(m)}$, (39) can be rewritten in DLP form in a similar way to the time domain algorithm as

$$\mathbf{x}_{n,f}^{(1)} = \bar{\mathbf{c}}_f^{*T} \bar{\mathbf{x}}_{n-D',f} + \mathbf{d}_{n,f} \quad (40)$$

$$\mathbf{d}_{n,f} = \sum_{k=0}^{D'-1} \left(\mathbf{h}_{k,f}^{(1)}\right)^{*T} \mathbf{s}_{n-k,f}. \quad (41)$$

where

$$\bar{\mathbf{c}}_f = \left[\left(\bar{\mathbf{c}}_f^{(1)}\right)^T, \left(\bar{\mathbf{c}}_f^{(2)}\right)^T\right]^T, \quad (2L_{c'} \times 1)$$

$$\bar{\mathbf{c}}_f^{(m)} = \left[\left(\mathbf{c}_{1,f}^{(m)}\right)^T, \ldots, \left(\mathbf{c}_{L_{c'},f}^{(m)}\right)^T\right]^T, \quad (L_{c'} \times 1)$$

TABLE I
SUMMARY OF DEREVERBERATION METHODS REGARDING MODEL OF SOURCE VARIANCE (TIME–VARYING OR STATIONARY), IMPLEMENTATION DOMAIN (TIME OR TIME–FREQUENCY), AND PREDICTION METHOD (WITH OR WITHOUT PREDICTION DELAY)

	FD-NDLP	TD-NDLP	DLP	LP
Variance	Time-varying		Stationary	Time-varying
Domain	Time-frequency	Time		
Prediction	w/ delay			w/o delay

and $L_{c'} = \lceil FL_c/K \rceil$ and $D' = \lceil FD/K \rceil$ are the order of the regression coefficients and the prediction delay in the time–frequency domain that are determined based on the corresponding values, L_c and D, in the time domain. Note that the index of the channel to be dereverberated, such as m in $\bar{c}^{(m',m)}$ in the time domain, is omitted in the time–frequency domain on the assumption that it is fixed at one.

C. Definition of Speech Model and Likelihood Function

With subband decomposition, the observation model is defined separately in individual subbands based on (40). Similarly, the desired signal is assumed to be a time-varying Gaussian process defined separately in individual subbands. According to these assumptions, the log likelihood function is also defined separately in individual subbands, and the dereverberation procedure can be accomplished separately in individual subbands.

Similar to the time domain algorithm, we introduce the pdf of the desired signal as the model of the speech signal. It is defined as

$$p(\mathbf{d}_{n,f}) = \mathcal{N}_c\left(\mathbf{d}_{n,f}; 0, \rho^2_{n,f}\right)$$

where $\mathcal{N}_c(\cdot)$ is the pdf of a complex Gaussian random process. With this definition, $\mathbf{d}_{n,f}$ is assumed to be a complex Gaussian process with a zero mean and a variance $\rho^2_{n,f} = E\{\mathbf{d}_{n,f}\mathbf{d}^*_{n,f}\}$. Because $\rho^2_{n,f}$ is assumed to take different values over time–frequency points, this model can precisely represent the characteristics of any time-varying power spectrum. Similar to the time domain algorithm, $\rho^2_{n,f}$ is not given in advance, and is taken as a parameter to be estimated.

Let $\rho^2_f = \{\rho^2_{1,f}, \rho^2_{2,f}, \ldots\}$ be a time series of $\rho^2_{n,f}$ for all frames n at a subband f, and $\theta_f = \{\bar{\mathbf{c}}_f, \rho^2_f\}$ be the parameter set to be estimated. Then, the log likelihood function for the dereverberation can be derived as follows:

$$\begin{aligned}\mathcal{L}_f(\theta_f) &= \sum_n \log p\left(\mathbf{d}_{n,f} = \mathbf{x}^{(1)}_{n,f} - \bar{\mathbf{c}}^{*T}_f \bar{\mathbf{x}}_{n-D',f}; \theta_f\right) \\ &= -\sum_n \frac{\left|\mathbf{x}^{(1)}_{n,f} - \bar{\mathbf{c}}^{*T}_f \bar{\mathbf{x}}_{n-D',f}\right|^2}{\rho^2_{n,f}} \\ &\quad - \sum_n \log \rho^2_{n,f} + \text{const.} \end{aligned} \tag{42}$$

D. Solution Based on Likelihood Maximization

The iterative optimization algorithm for maximizing (42) can be derived in the same way as the time domain algorithm, and is summarized as follows.

1) Initialize $\hat{\rho}^2_{n,f}$ as

$$\hat{\rho}^2_{n,f} = \max\{|\mathbf{x}_{n,f}|^2, \epsilon_f\}$$

where $\epsilon_f > 0$ is a lower bound of $\hat{\rho}^2_{n,f}$ at f.

2) Repeat the following until convergence.
   a) Update $\hat{\bar{\mathbf{c}}}_f$ as follows:

$$\hat{\bar{\mathbf{c}}}_f = \left(\sum_n \frac{\bar{\mathbf{x}}_{n-D',f}\bar{\mathbf{x}}^{*T}_{n-D',f}}{\rho^2_{n,f}}\right)^+ \sum_n \frac{\bar{\mathbf{x}}_{n-D',f}\mathbf{x}^{(1)*}_{n,f}}{\rho^2_{n,f}}.$$

   b) Update $\hat{\mathbf{d}}_{n,f}$ as $\hat{\mathbf{d}}_{n,f} = \mathbf{x}^{(1)}_{n,f} - \hat{\bar{\mathbf{c}}}^{*T}_f \bar{\mathbf{x}}_{n-D',f}$.
   c) Update $\hat{\rho}^2_{n,f}$ as follows:

$$\hat{\rho}^2_{n,f} = \max\{|\hat{\mathbf{d}}_{n,f}|^2, \epsilon_f\}.$$

The overall flow of the subband processing is shown in Fig. 3.

Even with the above algorithm, we need to calculate the variance-normalized covariance matrices at step 2a). However, compared with the time domain algorithm, the order of the regression coefficients can be substantially reduced from L_c to FL_c/K, and thus the matrix size can be much smaller. As a result, we can greatly reduce the computational cost.

The above time–frequency domain NDLP algorithm can also be shown to be robust with respect to the noise and modeling error terms in a way similar to the time-domain algorithm. Therefore, even without taking particular care with regard to such errors, the time–frequency domain algorithm can work effectively for dereverberation when the error level is relatively small. On the other hand, with more substantial errors, we can integrate the proposed dereverberation method with time–frequency domain noise reduction and/or blind source separation algorithms based on the likelihood maximization approach in a rather straightforward way [24], [25]. A discussion of such integration is beyond the scope of this paper.

V. EXPERIMENTS

This section describes our experimental investigation of the behavior of the proposed dereverberation method, NDLP, in both the time and time–frequency domains. We compare them with two existing methods particularly focusing on the computational efficiency (Exp1 in Section V-D), dereverberation accuracy (Exp2 in Section V-E), and robustness as regards acoustic noise (Exp3 in Section V-F), and model mismatch (Exp4 in Section V-G).

A. Methods to be Compared

We examined the behavior of the following four dereverberation methods, which we refer to as FD-NDLP, TD-NDLP, DLP, and LP in this paper. The differences between the methods are described below and also summarized in Table I.

1) FD-NDLP: This is the proposed dereverberation method described in Section IV-D. It is based on variance-normalized delayed linear prediction (NDLP) and works in the time–frequency domain.

2) TD-NDLP: This is the other proposed method described in Section II-D. It is also based on NDLP and works in the time

domain. By comparing TD-NDLP and FD-NDLP, we can confirm the effect of the time–frequency domain implementation.

3) DLP: This is an existing dereverberation method proposed in [20]. It is based on delayed linear prediction (DLP) without variance normalization and works in the time domain. When we set σ_t^2 so that it is constant and adopt the correlation method for the least square solution to the linear prediction analysis with TD-NDLP, DLP, and TD-NDLP are identical. So, we use DLP mainly to confirm the effect of variance normalization. With the original DLP, the dereverberation is performed as follows.

- First, regression coefficients are estimated from the observed signal, and the late reverberation is estimated by convolving the estimated regression coefficients with the observed signal. Both are performed in the time domain.
- Then, the energy of the late reverberation is reduced from the observed signal in the power spectral domain by spectral subtraction.

The above second step is introduced to make the dereverberation robust as regards the errors in the estimated regression coefficients. In this paper, however, to confirm the effect of the variance normalization in detail, we also examined the behavior of DLP when the estimated late reverberation was subtracted in the time domain.

4) LP: This is an existing method described as TVAR in [22]. It is based on linear prediction with no prediction delay. To appropriately deal with the short time correlation of the speech signal, the short time correlation matrices of the speech signal are estimated from the observed signal, and used for the dereverberation. We use LP mainly to confirm the effect of the introduction of the prediction delay D by comparing it with the other methods.

Note that we adopted pre-whitening as a pre-processing technique for TD-NDLP and DLP when calculating the regression coefficients as in [20].

B. Test Data Sets

As speech data for the experiments, we adopted city name utterances in the Japan Electronic Industry Development Association's Japanese Common Speech Data Corpus (JEIDA/JCSD). We extracted ten city name utterances from five female and five male speakers (a total of 100 city name utterances). Using these utterances, we prepared four test data sets, referred to as U1, U2, U5, and U10. Each data set contained a set of test utterances, and each test utterance contained concatenated city name utterances delivered by one speaker. The numbers of city name utterances included in a test utterance were 1, 2, 5, and 10 for U1, U2, U5, and U10, respectively, and the average lengths of the test utterances were 1.0, 2.0, 5.0, and 9.9 s for U1, U2, U5, and U10, respectively. The numbers of test utterances included in U1, U2, U5, and U10 were 100, 50, 20, and 10, respectively. In the experiments, dereverberation was performed for each test utterance in each data set, and the average performance of each dereverberation method was calculated for all the utterances in the data set.

The observed reverberant signals were synthesized by convolving these test utterances with two-channel RIRs measured in a reverberant room. We prepared four sets of RIRs, where the reverberation times of the room (R_{60}) were 0.1, 0.2, 0.5, and 1.0 s, respectively, and the Deutlichkeit values (D_{50}) were 0.98, 0.93, 0.83, and 0.73, respectively. The lengths of the RIRs were set at 800, 2400, 4000, and 8000 taps. An RIR with R_{60} =0.5 s was used for all the experiments, while the other RIRs were used only for Exp4 in Section V-G. For Exp3 in Section V-F, stationary white Gaussian noise was also added to the observed signal, where two different settings were adopted for the reverberant signal to noise ratio (RSNR), namely 30 and 10 dB.

C. Analysis Conditions

In this paper, all the experiments were conducted under the same analysis conditions described in the following. We set the sampling rate for all the signals at 8 kHz. With FD-NDLP, the number of subbands (including negative frequencies) and the over-sampling rate were set at 256 and 2, respectively. The order of the regression coefficients in each subband was determined on the assumption that the reverberation time decreases as the frequency increases. Concretely, for frequency ranges of 0–800, 800–1500, 1500–3000, and 3000–4000 Hz, the orders of the coefficients were set at $L_{c'} = 25, 20, 15$, and 10, respectively, which correspond to time domain filters with lengths of 0.4, 0.32, 0.24, and 0.16 s. The prediction delay was set at $D' = 2$, which corresponds to 32 ms in the time domain. With TD-NDLP, DLP, and LP, the order of the regression coefficients was set at $L_c = 3200$ (=0.4 s) and the prediction delay was set at $D = 32$ ms. With TD-NDLP and LP, the length of the short time frame was set at $L_f = 256$ (=32 ms). The maximum number of iterations for FD-NDLP, TD-NDLP, and LP was set at 1 in all cases.

D. Exp1: Computational Efficiency of Dereverberation

The computational efficiency of each dereverberation method was evaluated in terms of the real time factor (RTF). The RTF is defined as the ratio of the computing time required for the dereverberation processing to the time duration of the observed signal. When the RTF is less than one, it means that the dereverberation method can process a signal within a time shorter than the signal duration. The dereverberation methods were all implemented with MATLAB, and their computing times were measured by a MATLAB interpreter on a Linux computer with an Intel Pentium 4 CPU (3.6 GHz).

The average RTF of each dereverberation method was measured over all the test utterances in all the test data sets with a reverberation time of $R_{60} = 0.5$ s. Fig. 4 shows the resultant RTFs. By comparing the RTF of FD-NDLP with the RTFs of TD-NDLP and LP in the figure, we can confirm that the time–frequency implementation was capable of greatly improving the computational efficiency of the dereverberation based on the time-varying source models. In contrast, the RTFs of FD-NDLP and DLP were almost the same, and smaller than one. Although DLP is also time-domain algorithm, it can be implemented in a computationally efficient manner using the well-known Levinson–Durbin algorithm because the covariance matrix of the observed signal can be approximated by a correlation matrix according to the stationary source variance assumption.

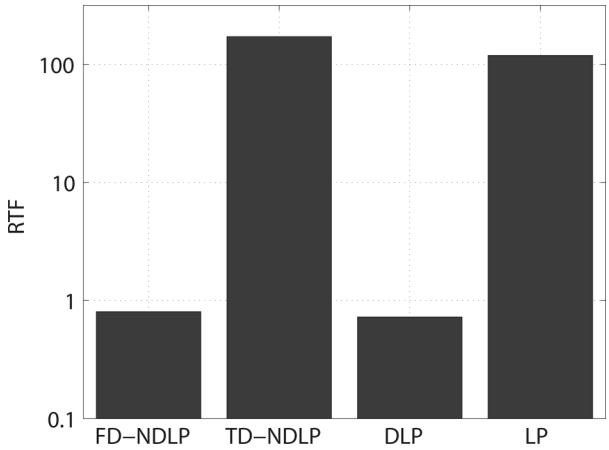

Fig. 4. Average real time factor for each dereverberation method over all test data sets. The vertical axis is represented using a logarithmic scale.

E. Exp2: Accuracy of Dereverberation

The dereverberation accuracy was evaluated in terms of average cepstral distortion (CD) over short time frames [29]. CD was originally defined as a measure of the distance between two log power spectra, and is one of the most frequently used objective measures for evaluating speech quality in the speech analysis area. The cepstral distortion in dB at a short time frame is defined as

$$\mathrm{CD} = (10/\ln 10)\sqrt{(\hat{\beta}_0 - \beta_0)^2 + 2\sum_{k=1}^{12}(\hat{\beta}_k - \beta_k)^2}$$

where $\hat{\beta}_k$ and β_k are the cepstral coefficients of the speech signal under evaluation and the original clean speech signal, respectively. k in β_k specifies the order of the cepstral coefficient to be evaluated. The zeroth-order cepstral coefficient represents the average of a log power spectrum over frequency and is referred to as an energy term hereafter, while the 1st to 12th order coefficients represent the envelope of the log power spectrum. The CD for clean speech signals is 0 dB by definition. The accuracy of each dereverberation method was evaluated based on the average CD over short time frames included in all the test utterances in each test data set. To ensure that the CD measure properly evaluated the accuracy of the late reverberation reduction by disregarding the effect of early reverberation, we applied cepstral mean normalization (CMN) to both the speech signal under evaluation and the original clean speech signal. After CMN, the average cepstral coefficients all become zero and thus we can mainly evaluate the time variation of the coefficients by the CD measure. Furthermore, we introduced an additional assumption with respect to the dereverberation errors at a non-speech frame. That is, when the energy term value of the speech signal under evaluation was less than that of the clean speech signal at a non-speech frame, we assumed that the reverberation energy was sufficiently reduced at this frame, and thus the CD measure was set at zero. A frame was defined as a non-speech frame when the level at this frame was less than -30 dB relative to the average level of an entire signal in the original clean speech data.

The results obtained with a reverberation time of $R_{60} = 0.5$ s are shown in Fig. 5(a). In the figure, the results for FD-NDLP, TD-NDLP, and LP are comparable and they reduced the CDs for all the data sets more effectively than DLP. We consider the increase in the CD for signals processed by DLP to be due to the increase in the reverberation energy where the observed signal has relatively little energy as discussed in Section III-B2. This can also be confirmed in the spectrograms shown in Section V-H. This result shows the effectiveness of the variance normalization for dereverberation.

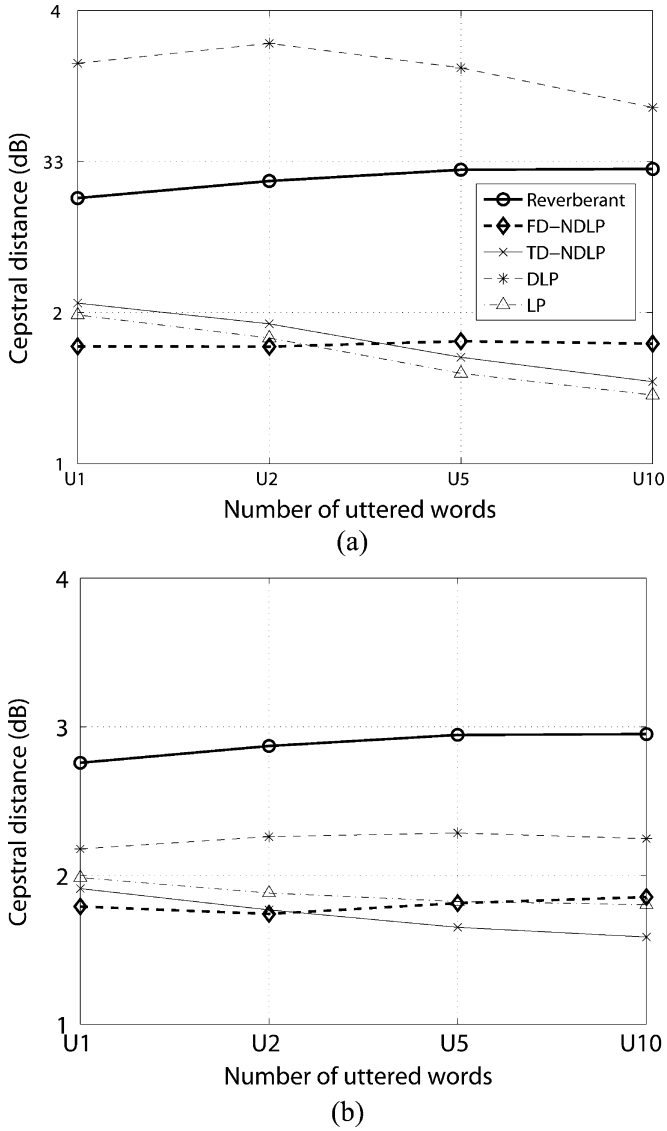

Fig. 5. Average cepstral distortions (CD) of the dereverberation methods when the dereverberation was performed by linear filtering (upper panel) and by spectral subtraction (lower panel). (a) Dereverberation by linear filtering. (b) Dereverberation by spectral subtraction.

When we compared the accuracy of the dereverberation methods in more detail, FD-NDLP was the best when the observation was short, namely for U1 and U2, while TD-NDLP and LP gradually improved and outperformed FD-NDLP as the observation became longer. We consider that the accuracy of FD-NDLP is limited due to the modeling errors caused by the approximated convolution in the time–frequency domain, and thus the estimation accuracy does not increase greatly as the observation becomes longer.

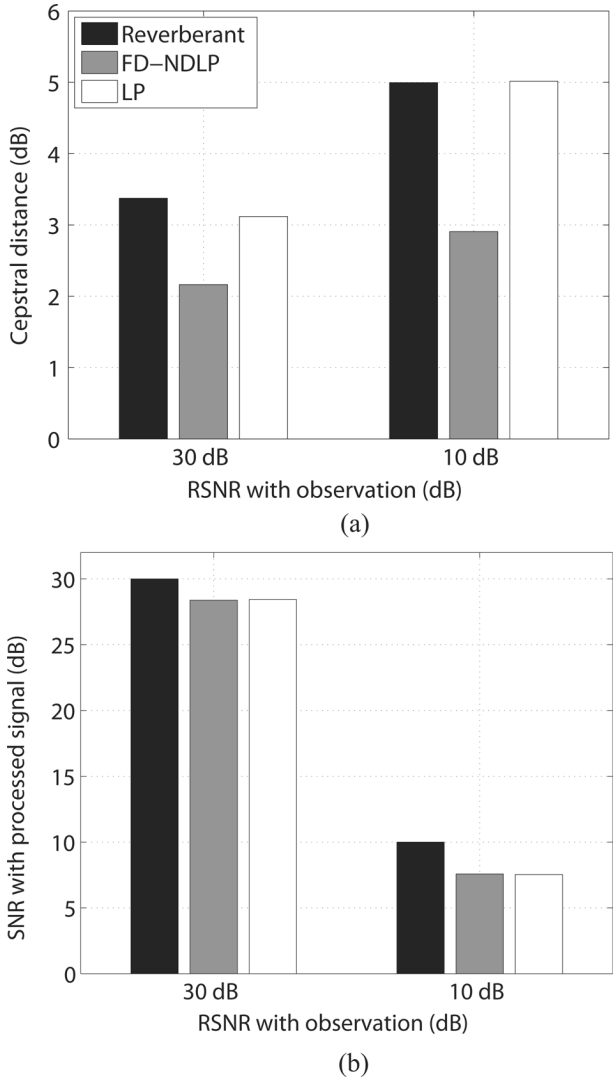

Fig. 6. Average cepstral distortions (CD) and signal-to-noise ratio (SNR) of dereverberated signals obtained by FD-NDLP and LP using U5. (a) CD. (b) SNR.

Fig. 5(b) shows the average CDs achieved by the dereverberation methods when the dereverberation was performed based not on linear filtering but on spectral subtraction, which was originally adopted for DLP in [20] to make the dereverberation robust against the errors in the estimated regression coefficients (see Section V-A3 for the dereverberation processing steps with the original DLP). The figure shows that the DLP accuracy was greatly improved by the introduction of spectral subtraction, and became close to that obtained with the other methods.

F. Exp3: Robustness Against Noise

To test the noise robustness of each dereverberation method, we evaluated the effect of the dereverberation on noisy reverberant speech. First, we represent a noisy reverberant signal as $x_t = z_t + b_t$, where z_t is a noise-free reverberant signal and b_t is the noise. Because our dereverberation approach is based on linear filtering, we can evaluate the effect of the dereverberation on the speech and the noise separately by applying the dereverberation filter separately to z_t and b_t after estimating the filter

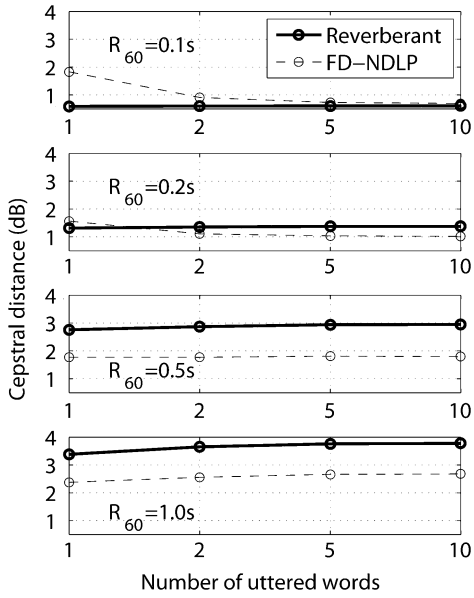

Fig. 7. Average CDs obtained by FD-NDLP under different R_{60} conditions. The order of the regression coefficients was fixed under all the conditions.

from noisy reverberant signal x_t. The resultant signals are denoted by $\check{z}_t$ and $\check{b}_t$. We evaluated the dereverberation effect on the speech from the CD of $\check{z}_t$, and we evaluated the dereverberation effect on the noise from the signal to noise ratio (SNR) of the dereverberated signal, which is defined as

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_t \check{z}_t^2}{\sum_t \check{b}_t^2}.$$

This SNR is used to confirm that the dereverberation filter does not amplify the noise too much.

Fig. 6 shows the CDs and SNRs of the dereverberated signals obtained by FD-NDLP and LP when stationary white Gaussian noise was added to the reverberant signals with RSNRs of 10 and 30 dB. The reverberation time was $R_{60} = 0.5$ s. Fig. 6(a) indicates that FD-NDLP effectively reduced the average CDs much more effectively than LP. By contrast, from Fig. 6(b), we can confirm that both methods slightly increased the noise level, but the increase was not very large.

These results suggest that the introduction of the prediction delay greatly improved the noise robustness of the dereverberation method.

G. Exp4: Robustness Against Model Mismatch

We also evaluated the accuracy of the dereverberation by FD-NDLP under different reverberation time conditions, namely $R_{60} = 0.1, 0.2, 0.5$, and 1.0 s without modifying the orders of the regression coefficients. In theory, the order of the regression coefficients should be approximately equal to the channel order of the observation system when using two microphones. So, there are certain channel order mismatches between the observation system and the model of the dereverberation in this experiment. For example, the order of the regression coefficients was too large for $R_{60} = 0.1$ and 0.2 s while it was too small for $R_{60} = 1.0$ s.

As shown in Fig. 7, the average CDs improved under conditions of $R_{60} = 0.5$ and 1.0 s and under a condition where

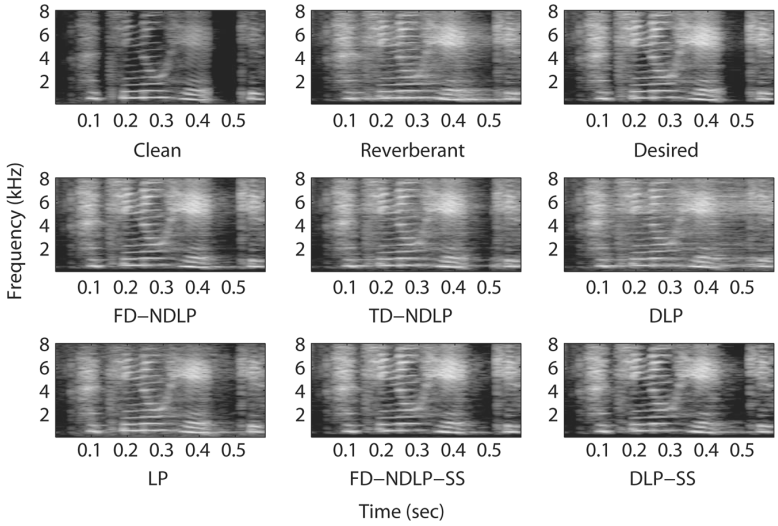

Fig. 8. Spectrograms of clean (top left), reverberated (top middle, $R_{60} = 0.5$ s), and desired (top right) signals of a female utterance in U2, signals dereverberated based on linear filtering with FD-NDLP (center left), TD-NDLP (center middle), DLP (center right), and LP (bottom left), and signals dereverberated based on spectral subtraction (SS) with FD-NDLP (bottom middle), and DLP (bottom right).

$R_{60} = 0.2$ s and the length of the observation was equal to or longer than U2. Although the quality of the signal for $R_{60} = 0.1$ s was degraded by the dereverberation, this degradation could be disregarded as the observation became sufficiently long. This result shows the robustness of the proposed method, FD-NDLP, against a mismatch of the order of the regression coefficients.

H. Example Spectrograms

Example spectrograms of clean, reverberated, desired, and dereverberated signals corresponding to a female utterance in U2 are shown in Fig. 8. From the figure, we can confirm that the reverberation energy was effectively reduced based on linear filtering (namely, without spectral subtraction) with FD-NDLP, TD-NDLP, and LP, where the frequency structure of the desired signal was clearly recovered. In contrast, with DLP, the dereverberation was effective only when the dereverberation was performed based on spectral subtraction.

VI. Concluding Remarks

This paper proposed a statistical model-based speech dereverberation approach, referred to as a variance-normalized delayed linear prediction (NDLP). With this approach, thanks to the framework of delayed linear prediction (DLP), an inverse system that can cancel out the effect of late reverberation can be robustly estimated even in the presence of noise. In addition, owing to the use of variance normalization, NDLP can improve the dereverberation accuracy especially with relatively short (e.g., a few seconds) observations in comparison with DLP. Furthermore, it is straightforward to implement NDLP in a computationally efficient manner in the time–frequency domain. The behavior of NDLP was investigated in detail by both mathematical analysis and experiments, and as a result, NDLP is shown to be very effective and computationally efficient in achieving dereverberation even in the presence of background noise compared with existing dereverberation approaches.

Future work will include the elaboration of the model to deal with the time-varying nature of the speech signal for more precise dereverberation. Real-time processing is also an important issue for many real applications.

Appendix A
Derivation of Observation Model With Delayed Linear Prediction

We set $\bar{c}^{(m)} = H(H^T H)^{-1}\bar{h}_l^{(m)}$ where H is the convolution matrix defined by (21) and $\bar{h}_l^{(m)}$ is the early room impulse response defined as (22). Note that $H^T H$ is non-singular when H is assumed to be full column rank. Then, the right-hand side of (6) can be rewritten as

$$\begin{aligned} x_t^{(m)} &= r_t^{(m)} + d_t^{(m)} \\ &= \left(\bar{h}_l^{(m)}\right)^T \bar{s}_{t-D} + d_t^{(m)} \\ &= \left(\bar{h}_l^{(m)}\right)^T (H^T H)^{-1} H^T H \bar{s}_{t-D} + d_t^{(m)} \\ &= \left(\bar{c}^{(m)}\right)^T \bar{x}_{t-D} + d_t^{(m)} \end{aligned}$$

where we use (3) and (20). This means that we can find $\bar{c}^{(m)}$ that satisfies (6), or equivalently (5), given any invertible room impulse response.

Appendix B
Derivation of (9)

Equation (8) can be rewritten as

$$\begin{aligned} \mathcal{L}(\theta) &= \log p(\bar{x}_{\langle \mathcal{T}, -\infty \rangle}; \theta) \\ &= \log p\left(\bar{x}^{(2)}_{\langle \mathcal{T}, \mathcal{T}-D+1 \rangle}, \bar{x}^{(1)}_{\langle \mathcal{T}, \mathcal{T}-D+1 \rangle} \bar{x}_{\langle \mathcal{T}-D, -\infty \rangle}; \theta\right) \\ &= \log p\left(\bar{x}^{(2)}_{\langle \mathcal{T}, \mathcal{T}-D+1 \rangle} \middle| \bar{x}^{(1)}_{\langle \mathcal{T}, \mathcal{T}-D+1 \rangle}, \bar{x}_{\langle \mathcal{T}-D, -\infty \rangle}; \theta\right) \\ &\quad + \log p\left(\bar{x}^{(1)}_{\langle \mathcal{T}, \mathcal{T}-D+1 \rangle}, \bar{x}_{\langle \mathcal{T}-D, -\infty \rangle}; \theta\right). \end{aligned}$$

The second term of the last equation above can be rewritten by repeatedly expanding a joint pdf into two conditional pdfs and a prior joint pdf as

$$\begin{aligned}
&\log p\left(\bar{x}^{(1)}_{\langle \mathcal{T},\mathcal{T}-D+1\rangle}, \bar{x}_{\langle \mathcal{T}-D,-\infty\rangle}; \theta\right) \\
&\quad = \log p\left(x^{(2)}_{\mathcal{T}-D} \middle| \bar{x}^{(1)}_{\langle \mathcal{T},\mathcal{T}-D\rangle}, \bar{x}_{\langle \mathcal{T}-D-1,-\infty\rangle}; \theta\right) \\
&\qquad + \log p\left(x^{(1)}_{\mathcal{T}} \middle| \bar{x}^{(1)}_{\langle \mathcal{T}-1,\mathcal{T}-D\rangle}, \bar{x}_{\langle \mathcal{T}-D-1,-\infty\rangle}; \theta\right) \\
&\qquad + \log p\left(\bar{x}^{(1)}_{\langle \mathcal{T}-1,\mathcal{T}-D\rangle}, \bar{x}_{\langle \mathcal{T}-D-1,-\infty\rangle}; \theta\right) \\
&\quad = \sum_{t=1}^{\mathcal{T}} \log p\left(x^{(2)}_{t-D} \middle| \bar{x}^{(1)}_{\langle t,t-D\rangle}, \bar{x}_{\langle t-D-1,-\infty\rangle}; \theta\right) \\
&\qquad + \sum_{t=1}^{\mathcal{T}} \log p\left(x^{(1)}_{t} \middle| \bar{x}^{(1)}_{\langle t-1,t-D\rangle}, \bar{x}_{\langle t-D-1,-\infty\rangle}; \theta\right) \\
&\qquad + \log p\left(\bar{x}^{(1)}_{\langle 0,1-D\rangle}, \bar{x}_{\langle -D,-\infty\rangle}; \theta\right).
\end{aligned}$$

Then, (8) can be rewritten as

$$\begin{aligned}
\mathcal{L}(\theta) &= \log p\left(\bar{x}^{(2)}_{\langle \mathcal{T},\mathcal{T}-D+1\rangle} \middle| \bar{x}^{(1)}_{\langle \mathcal{T},\mathcal{T}-D+1\rangle}, \bar{x}_{\langle \mathcal{T}-D,-\infty\rangle}; \theta\right) \\
&\quad + \sum_{t=1}^{\mathcal{T}} \log p\left(x^{(2)}_{t-D} \middle| \bar{x}^{(1)}_{\langle t,t-D\rangle}, \bar{x}_{\langle t-D-1,-\infty\rangle}; \theta\right) \\
&\quad + \sum_{t=1}^{\mathcal{T}} \log p\left(x^{(1)}_{t} \middle| \bar{x}^{(1)}_{\langle t-1,t-D\rangle}, \bar{x}_{\langle t-D-1,-\infty\rangle}; \theta\right) \\
&\quad + \log p\left(\bar{x}^{(1)}_{\langle 0,1-D\rangle}, \bar{x}_{\langle -D,-\infty\rangle}; \theta\right). \qquad (43)
\end{aligned}$$

In the following, we disregard the first and the fourth terms on the right-hand side of the above equation assuming $\mathcal{T}$ to be sufficiently large and $x^{(m)}_t = 0$ for $t \leq 0$, because the information on the observation increases only in the second and third terms as $\mathcal{T}$ becomes large. In addition, we consider that $x^{(2)}_{t-D}$ can be uniquely determined given $\bar{x}^{(1)}_{\langle t,t-D\rangle}$ and $\bar{x}_{\langle t-D-1,-\infty\rangle}$, and thus we disregard the second term (see Appendix A of [22]). Furthermore, based on the observation model (6) and given $\bar{x}_{t-D}$, the probabilistic uncertainty of $x^{(1)}_t$ in the third term is derived only from that of $d^{(1)}_t$ and $x^{(1)}_t$ does not depend on $\bar{x}^{(1)}_{\langle t-1,t-D\rangle}$ and $\bar{x}^{(1)}_{\langle t-D-1,-\infty\rangle}$. As a result, the log likelihood function can be rewritten as (9).

Appendix C
Derivation of (26)

It is easy to rewrite (25) in the following quadratic form:

$$\begin{aligned}
\mathcal{L}_1(\theta) = \left(H^T\bar{c}^{(1)} - \left(\bar{h}^{(1)}_l + \bar{\xi}\right)\right)^T \\
\times\Gamma\left(H^T\bar{c}^{(1)} - \left(\bar{h}^{(1)}_l + \bar{\xi}\right)\right) + \text{const}
\end{aligned}$$

where $\Gamma = \sum_{t=1}^{\mathcal{T}} \bar{s}_{t-D}\bar{s}^T_{t-D}/\hat{\sigma}^2_t$. It is obvious that (26) is the minimum-norm solution that maximizes (25).

References

[1] J. L. Flanagan, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, no. 11, pp. 1508–1518, 1985.

[2] G. W. Elko, "Superdirective microphone arrays," in *Acoustic Signal Processing for Telecommunication*, S. Gay and J. Benesty, Eds. Norwell, MA: Kluwer, 2000, pp. 181–235.

[3] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "A method based on the MTF concept for dereverberating the power envelope from the reverberant signal," in *Proc. 2003 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, vol. 1, pp. 840–843.

[4] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. ICASSP'05*, 2005, pp. 173–176.

[5] T. Yoshioka, H. Kameoka, T. Nakatani, and H. G. Okuno, "Statistical models for speech dereverberation," in *Proc. WASPAA'09*, 2009, pp. 145–148.

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

[7] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. APS*, vol. 2007, 2007, Article-ID 34013.

[8] S. C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Commun.*, vol. 39, pp. 65–78, 2003.

[9] B. W. Gillespie, H. S. Malvar, and D. A. F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP'01)*, 2001, vol. 6, pp. 3701–3704.

[10] K. Furuya and A. Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1579–1591, Jul. 2007.

[11] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. IWAENC'05*, 2005.

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

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

[14] Y. Huang, J. Benesty, and J. Chen, "Speech acquisition and enhancement in a reverberant, cocktail-party-like environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'06)*, 2006, vol. V, pp. 25–28.

[15] D. T. M. Slock, "Blind fractionally-spaced equalization, perfectreconstruction filter-banks and multichannel linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'94)*, 1994, vol. 4, pp. 585–588.

[16] K. Abed-Meraim, E. Mouline, and P. Loubation, "Prediction error method for second-order blind identification," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 694–705, Mar. 1997.

[17] D. Gesbert and P. Duhamel, "Robust blind channel identification and equalization based on multi-step predictors," in *Proc. ICASSP'97*, 1997, pp. 3621–3624.

[18] M. Miyoshi, "Estimating AR parameter-sets for linear-recurrent signals in convolutive mixtures," in *Proc. ICA'03*, 2003, pp. 585–589.

[19] M. Triki and D. T. M. Slock, "Delay and predict equalization for blind speech dereverberation," in *Proc. ICASSP'05*, 2005, vol. V, pp. 97–100.

[20] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 534–545, May 2009.

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

[22] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, Nov. 2008.

[23] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. ICASSP'08*, 2008, pp. 85–88.

[24] T. Yoshioka, T. Nakatani, and M. Miyoshi, "An integrated method for blind separation and dereverberation of convolutive audio mixtures," in *Proc. EUSIPCO'08*, 2008.
[25] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.
[26] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Incremental estimation of reverberation with uncertainty using prior knowledge of room acoustics for speech dereverberation," in *Proc. IWAENC'08*, 2008.
[27] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. ICASSP'09*, 2009, pp. 3733–3736.
[28] S. Weiss and R. W. Stewart, "Fast implementation of oversampled modulated filter banks," *IEE Electron. Lett.*, vol. 36, no. 17, pp. 1502–1503, 2000.
[29] S. Furui, *Digital Speech Processing, Synthesis, and Recognition, Second Edition, Revised and Expanded*. New York: Marcel Dekker, 2001.

Tomohiro Nakatani (M'03–SM'06) received the B.E., M.E., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 1989, 1991, and 2002, respectively.

He is a Senior Research Scientist (Supervisor) of NTT Communication Science Labs, NTT Corporation, Kyoto. Since he joined NTT Corporation as a Researcher in 1991, he has been investigating speech enhancement technologies for developing intelligent human–machine interfaces. Since 2005, he has visited the Georgia Institute of Technology, Atlanta, as a Visiting Scholar for a year where he worked with Prof. B.-H. Juang. Since 2008, he has been a Visiting Assistant Professor in the Department of Media Science, Nagoya University, Nagoya, Japan.

Dr. Nakatano was honored to receive the 1997 JSAI Conference Best Paper Award, the 2002 ASJ Poster Award, the 2005 IEICE Best Paper Award, and the 2009 ASJ Technical Development Award. He has been a member of the IEEE Signal Processing Audio and Acoustics Technical Committee since 2009, a member of the IEEE CAS Blind Signal Processing Technical Committee since 2007, and an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING since 2008. He served as a Technical Program Chair of IEEE WASPAA-2007. He is a member of IEICE and ASJ.

Takuya Yoshioka (M'08) received the B.E., M.Inf., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 2004, 2006, and 2010, respectively.

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

Dr. Yoshioka is a member of ASJ.

Keisuke Kinoshita (M'05) received the M.Eng. and Ph.D. degrees from Sophia University, Tokyo, Japan, in 2003 and 2010, respectively.

He is currently a Researcher at NTT Communication Science Labs, Kyoto, Japan, and engaged in the research on speech and audio signal processing.

Dr. Kinoshita was honored to receive the 2006 IEICE Paper Awards and the 2009 ASJ Technical Development Awards. He is a member of ASJ and IEICE.

Masato Miyoshi (M'87–SM'04) received the M.E. and Ph.D. degrees from Doshisha University, Kyoto, Japan, in 1983 and 1991, respectively.

From 1983 to 2009, he engaged, as a research staff member of Nippon Telegraph and Telephone Corp., in the research on signal processing theory and its application to acoustic technology. Since 2009, he has been with the Graduate School of Natural Science and Technology, Kanazawa University, Ishikawa, Japan.

Dr. Miyoshi was honored to receive the 1988 IEEE Senior Awards, the 1989 ASJ Awaya Prize Young Researcher Award, the 1990 and 2006 ASJ Sato Prize Paper Awards, the 2005 IEICE Best Paper Award, and the 2009 ASJ Technical Development Award, respectively. He is a member of ASJ, IEICE, and AES.

Biing-Hwang (Fred) Juang (M'80–SM'87–F'91) received the Ph.D. degree from the University of California, Santa Barbara.

He had worked at Speech Communications Research Laboratory (SCRL) and Signal Technology, Inc. (STI) on a number of Government-sponsored research projects. Notable accomplishments during the period include development of vector quantization for voice applications, voice coders at extremely low bit rates, 800 bps, and around 300 bps, and robust vocoders for use in satellite communications. He was also a co-Principal Investigator for the project on co-channel separation of speech signals sponsored by the Department of Defense. He subsequently joined the Acoustics Research Department of Bell Laboratories, working in the area of speech enhancement, coding and recognition. He became Director of Acoustics and Speech Research at Bell Labs in 1996, and Director of Multimedia Technologies Research at Avaya Labs (a spin-off of Bell Labs) in 2001. His group continued the long heritage of Bell Labs in speech communication research, including, most notably, the invention of electret microphone, network echo canceller, a series of speech CODECs, and key algorithms for signal modeling and automatic speech recognition. In the past few years, he and his group developed a speech server for applications such as AT&T's advanced 800 calls and the Moviefone, the Perceptual Audio Coder (PAC) for digital audio broadcasting in North America (in both terrestrial and satellite systems), and a world-first real-time full-duplex hands-free stereo teleconferencing system. He has published extensively, including the book *Fundamentals of Speech Recognition* (Prentice-Hall, 1993), coauthored with L. R. Rabiner, and holds about 20 patents. He joined the Georgia Institute of Technology, Atlanta, in 2001 holding the Motorola Foundation Chair Professorship and is a Georgia Research Alliance Eminent Scholar.

Prof. Juang has served as Editor-in-Chief for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, and a number of positions in the IEEE Signal Processing Society, including Chair of its Fellow Evaluation Committee. He has received a number of technical awards, notable among which are several Best Paper awards in the area of speech communications and processing, the Technical Achievement Award from the Signal Processing Society of the IEEE, and the IEEE Third Millennium Medal. He is a Fellow of Bell Laboratories, a member of the U.S. National Academy of Engineering, and an Academician of Academia Sinica.