NOISE ROBUST SPEECH DEREVERBERATION WITH KALMAN SMOOTHER

Masahito Togami and Yohei Kawaguchi

Central Research Laboratory, Hitachi Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan

ABSTRACT

A speech dereverberation method is proposed that is robust against background noise. In contrast to conventional methods based on the linear prediction of the given microphone input signal, in which the linear prediction coefficients are not fully optimized when there is background noise, the proposed method optimizes the coefficients by linear prediction of the noiseless reverberant speech signal even when there is background noise. The noiseless reverberant speech signal and the parameters are iteratively updated on the basis of the expectation maximization algorithm. In the expectation step, sufficient statistics of latent variables which include noiseless reverberant speech signal are estimated using the Kalman smoother. Unlike the standard Kalman smoother, which uses a time-invariant covariance matrix as a state-transition covariance matrix, the proposed method utilizes a time-varying covariance matrix, enabling it to meet the time-varying speech characteristics. The parameters are updated so that the Q function is increased in the maximization step. Experimental results show that the proposed method is superior to conventional methods under noisy conditions.

Index Terms— Noise reduction, dereverberation, kalman smoother

1. INTRODUCTION

Room reverberation and background noise are major problems for speech recording and speech communication systems. Their levels must therefore be reduced simultaneously.

Several dereverberation techniques using multichannel inverse filtering have been developed [1][2][3][4][5]. A commonly used dereverberation technique is multi-step linear prediction (MSLP) [3], which reduces late reverberation by linear prediction of the microphone input signal. This technique is easy to implement and reduces reverberation effectively when there is no background noise. If the microphone input signal is contaminated by background noise, the linear prediction coefficients are updated so as to minimize the summation of the residual late reverberation and residual background noise after linear prediction. Therefore, if the background noise is dominant, dereverberation performance is degraded.

We have developed a speech dereverberation technique based on linear prediction of noiseless reverberant speech signal that overcomes this problem. The linear prediction coefficients are optimized so as to minimize only the amount of the residual reverberation after linear prediction even when the background noise is dominant. Since the noiseless reverberant speech signal is not given in advance, the proposed method estimates the noiseless reverberant speech signal and dereverberation parameters in an iterative manner using the expectation maximization (EM) algorithm [6]. In the expectation (E) step, the Kalman smoother [7] is used to obtain sufficient statistics of latent variables which include the noiseless reverberant speech signal. The state-transition noise is related to the direct path of the speech signal in the state-transition equation. By extracting the state-transition noise, we can obtain the noiseless and dereverberated speech signal. A time-varying covariance matrix is used as a state-transition covariance matrix to meet the time-varying characteristics of the speech sources. The dereverberation parameters are updated so as to increase the Q function in the maximization (M) step. Experimental results show that the proposed method can reduce reverberation and background noise effectively in noisy environments.

2. PROBLEM STATEMENT

2.1. Microphone input signal model

The focus here is on multichannel processing. Assuming that there is a single speech source and background noise, we can model the multichannel microphone input signal at frequency f and for frame τ as

$$\boldsymbol{x}_{f,\tau} = \sum_{l=0}^{L_1-1} s_{f,\tau-l}\boldsymbol{a}_{f,l} + \boldsymbol{w}_{f,\tau}, \tag{1}$$

where $\boldsymbol{x}_{f,\tau} = [\ x_{1,f,\tau} \quad \cdots \quad x_{N_m,f,\tau}\ ]^T$, T is the transpose operator of a matrix or vector, N_m is the number of microphones, L_1 is the tap length of the acoustic transfer function (ATF), $s_{f,\tau}$ is the source signal at each time-frequency point, $\boldsymbol{a}_{f,l}$ is the lth tap of the ATF of the speech source at each time-frequency point, and $\boldsymbol{w}_{f,\tau}$ is the background noise term (white Gaussian noise).

2.2. Conventional autoregressive model for microphone input signal

In the conventional methods [3], (1) is transformed into an autoregressive model of the microphone input signal:

$$\boldsymbol{x}_{f,\tau} = \sum_{l=0}^{D-1} s_{f,\tau-l}\boldsymbol{a}_{f,l} + \sum_{l=D}^{L_2-1} \boldsymbol{W}_{f,l}\boldsymbol{x}_{f,\tau-l} + \boldsymbol{b}_{f,\tau}, \tag{2}$$

where $L_2 = L_1 + L_i - 1$, L_i is the length of the inverse filter of the ATF, D is the tap-length of the early reflection, and $\boldsymbol{b}_{f,\tau}$ is the convolutive background noise term with the ATF of the speech source and the inverse filter.

In the conventional time-varying source-model-based dereverberation techniques, such as that of Nakatani et al. [9], the linear prediction coefficients $\boldsymbol{W}_{f,l}$ are estimated using

$$\hat{\boldsymbol{W}}_{f,l} = \underset{\boldsymbol{W}_{f,l}}{\operatorname{argmin}} \sum_{\tau=1}^{L_T} \mathcal{G}(\boldsymbol{W}_{f,l}, \boldsymbol{x}_{f,\tau}), \tag{3}$$

$$\mathcal{G}(\boldsymbol{W}_{f,l}, \boldsymbol{x}_{f,\tau}) = (\boldsymbol{x}_{f,\tau} - \sum_{l=D}^{L_2-1} \boldsymbol{W}_{f,l}\boldsymbol{x}_{f,\tau-l})^H \boldsymbol{R}_{x,f,\tau}^{-1} (\boldsymbol{x}_{f,\tau} - \sum_{l=D}^{L_2-1} \boldsymbol{W}_{f,l}\boldsymbol{x}_{f,\tau-l}), \tag{4}$$

where L_T is the number of the frames used for parameter optimization, H is the Hermite transpose operator of a matrix or vector, and $\boldsymbol{R}_{x,f,\tau} = E[\boldsymbol{x}_{f,\tau}\boldsymbol{x}_{f,\tau}^H]$. Let $\boldsymbol{x}_{f,\tau}$ be divided into speech term $\boldsymbol{c}_{f,\tau} = \sum_{l=0}^{L_1-1} s_{f,\tau-l}\boldsymbol{a}_{f,\tau,l}$ and background noise term $\boldsymbol{w}_{f,\tau}$:

$$\boldsymbol{x}_{f,\tau} = \boldsymbol{c}_{f,\tau} + \boldsymbol{w}_{f,\tau}. \tag{5}$$

Assuming that $\boldsymbol{c}_{f,\tau}$ and $\boldsymbol{w}_{f,\tau}$ are uncorrelated, we can approximate $\sum_{\tau=1}^{L_T} \mathcal{G}(\boldsymbol{W}_{f,l}, \boldsymbol{x}_{f,\tau})$:

$$\sum_{\tau=1}^{L_T} \mathcal{G}(\boldsymbol{W}_{f,l}, \boldsymbol{x}_{f,\tau}) \approx \sum_{\tau=1}^{L_T} \mathcal{G}(\boldsymbol{W}_{f,l}, \boldsymbol{c}_{f,\tau}) + \sum_{\tau=1}^{L_T} \mathcal{G}(\boldsymbol{W}_{f,l}, \boldsymbol{w}_{f,\tau}). \tag{6}$$

The first term is the amount of residual reverberation after linear prediction by $\boldsymbol{W}_{f,l}$, and the second term is the amount of residual background noise. The conventional method thus works by updating the linear prediction coefficients so as to minimize the summation of the residual reverberation and the residual background noise after linear prediction. If the background noise is small, the first term is dominant. Therefore, linear prediction coefficients that can reduce reverberation accurately should be obtained. However, if the background noise is dominant, dereverberation performance is degraded because the linear prediction coefficients are updated so as to minimize the residual background noise.

3. PROPOSED METHOD

3.1. Linear prediction with latent multichannel source signal

Our proposed speech dereverberation method avoids performance degradation due to background noise by minimizing only the residual reverberation after linear prediction instead of minimizing the summation of the residual reverberation and residual background noise. Since the noiseless speech signal cannot be obtained in advance, sufficient statistics of latent variables which include the noiseless speech signal are estimated by using the Kalman smoother [7]. The linear prediction coefficients are obtained by using the estimated statistics as follows:

$$\hat{\boldsymbol{W}}_{f,l} = \underset{\boldsymbol{W}_{f,l}}{\operatorname{argmin}} \sum_{\tau=1}^{L_T} E[\mathcal{G}(\boldsymbol{W}_{f,l}, \boldsymbol{c}_{f,\tau})], \tag{7}$$

where E is the operator for mathematical expectation. A block diagram of the proposed method is shown in Fig. 1. The Kalman smoother corresponds to the E step in the EM algorithm [6]. In the M step, the dereverberation parameters are updated using the statistics obtained in the E step so as to increase the Q function. The noiseless speech signal and dereverberation parameters are thereby updated in an iterative manner.

3.2. Sufficient statistics estimation

Sufficient statistics of latent variables which include the noiseless speech signal are the minimum mean square error (MMSE) estimate and the mean square error (MSE) of the latent variables. These statistics are effectively calculated by the Kalman smoother. At first, the microphone input signal is converted into a state-transition model and an observation model.

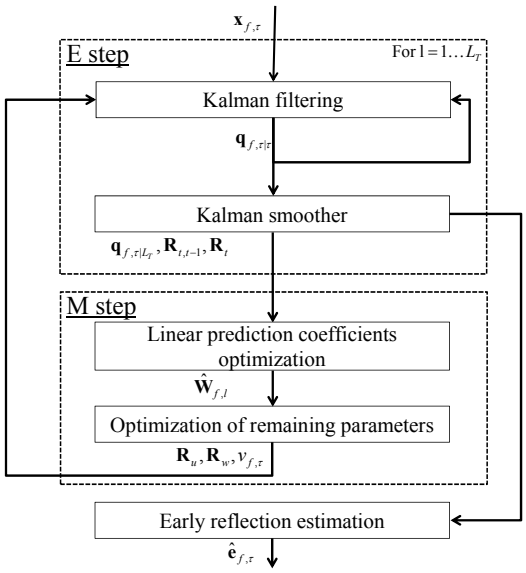

Fig. 1. Block diagram of proposed method

3.2.1. State-transition model

The noiseless reverberant speech signal in the microphone input signal can be expressed by the following state-transition model as

$$\boldsymbol{q}_{f,\tau} = \boldsymbol{A}_f \boldsymbol{q}_{f,\tau-1} + \boldsymbol{u}_{f,\tau}, \tag{8}$$

where $\boldsymbol{A}_f$ is the state-transition matrix, $\boldsymbol{q}_{f,\tau}$ is a state vector, $\boldsymbol{q}_{f,\tau} = [\ \boldsymbol{c}_{f,\tau}^H \quad \boldsymbol{c}_{f,\tau-1}^H \quad \boldsymbol{c}_{f,\tau-L_2+2}^H\ ]^H$, $\boldsymbol{A}_f$ is defined as

$$\boldsymbol{A}_f = \begin{bmatrix} \boldsymbol{0}_{N_m \times N_m(D-1)} & \boldsymbol{W}_{f,D} & \cdots & \boldsymbol{W}_{f,L_2-1} \\ \boldsymbol{I}_{N_m \times N_m} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0}_{N_m \times N_m} & \boldsymbol{I}_{N_m \times N_m} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0}_{N_m \times N_m} & \boldsymbol{0}_{N_m \times N_m} & \boldsymbol{I}_{N_m \times N_m} & \boldsymbol{0} \\ & \vdots & & \end{bmatrix}, \tag{9}$$

and $\boldsymbol{u}_{f,\tau}$ is defined as

$$\boldsymbol{u}_{f,\tau} = [\ \boldsymbol{e}_{f,\tau}^T \quad \boldsymbol{0} \quad \boldsymbol{0}\ ]^T, \tag{10}$$

where $\boldsymbol{e}_{f,\tau}$ is an $M \times M$ matrix. This matrix is the summation of the direct path and the early reflection term:

$$\boldsymbol{e}_{f,\tau} = \sum_{l=0}^{D-1} s_{f,\tau-l}\boldsymbol{a}_{f,l}. \tag{11}$$

The speech sources can be modeled as non-stationary source signals that are located at the same location while talking. Under these conditions, and similar to the modeling used in the conventional method [5], the probabilistic distribution of $\boldsymbol{e}_{f,\tau}$ is modeled as a multichannel time-varying Gaussian distribution [10] with a $\boldsymbol{0}$-mean vector:

$$p(\boldsymbol{e}_{f,\tau}) = \mathcal{N}(\boldsymbol{e}_{f,\tau}; \boldsymbol{0}, v_{f,\tau}\boldsymbol{R}_{u,f}), \tag{12}$$

where $v_{f,\tau}$ is a time-varying scalar coefficient which reflects the time-varying characteristics of the speech source and $\boldsymbol{R}_{u,f}$ is a time-invariant matrix which reflects the time-invariant characteristics of the speech source location.

3.2.2. Observation model

The microphone input signal which consists of the noiseless reverberant speech and the background noise can be derived as the following observation model:

$$\boldsymbol{x}_{f,\tau} = \boldsymbol{F}\boldsymbol{q}_{f,\tau} + \boldsymbol{w}_{f,\tau}, \tag{13}$$

where

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{I}_{N_m \times N_m} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \tag{14}$$

The probabilistic distribution of observation noise $\boldsymbol{w}_{f,\tau}$ is approximately modeled as a stationary Gaussian distribution with a $\boldsymbol{0}$-mean vector as follows:

$$p(\boldsymbol{w}_{f,\tau}) = \mathcal{N}(\boldsymbol{w}_{f,\tau}; \boldsymbol{0}, \boldsymbol{R}_{w,f}), \tag{15}$$

where $\boldsymbol{R}_{w,f}$ is the time-invariant covariance matrix of $\boldsymbol{w}_{f,\tau}$.

3.2.3. Kalman smoother for sufficient statistics estimation

The MMSE estimate and the MSE of the state vector $\boldsymbol{q}_{f,\tau}$ are observed under the condition that the microphone input signal $\boldsymbol{x}_{f,\tau}$ ($\tau = 1 \dots L_T$) is given.

Kalman filter:

$$\begin{aligned}
\boldsymbol{q}_{f,1|0} &= \boldsymbol{\pi}_f, && (16)\\
\boldsymbol{R}_{f,1|0} &= \boldsymbol{z}_f, && (17)\\
\boldsymbol{q}_{f,\tau|\tau-1} &= \boldsymbol{A}_f \boldsymbol{q}_{f,\tau-1|\tau-1}, && (18)\\
\boldsymbol{q}_{f,\tau|\tau} &= \boldsymbol{q}_{f,\tau|\tau-1} + \boldsymbol{K}_{f,\tau}(\boldsymbol{x}_{f,\tau} - \boldsymbol{F}\boldsymbol{q}_{f,\tau|\tau-1}), && (19)\\
\boldsymbol{K}_{f,\tau} &= \boldsymbol{R}_{f,\tau|\tau-1}\boldsymbol{F}^H \boldsymbol{R}^{-1}_{x,f,\tau}, && (20)\\
\boldsymbol{R}_{f,\tau|\tau-1} &= \boldsymbol{R}_{f,\tau-1|\tau-1} + v_{f,\tau}\boldsymbol{R}_{u,f}, && (21)\\
\boldsymbol{R}_{x,f,\tau} &= \boldsymbol{R}_{w,f} + \boldsymbol{F}\boldsymbol{R}_{f,\tau|\tau-1}\boldsymbol{F}^H, && (22)\\
\boldsymbol{R}_{f,\tau|\tau} &= (\boldsymbol{I} - \boldsymbol{K}_{f,\tau}\boldsymbol{F})(\boldsymbol{R}_{f,\tau|\tau-1}), && (23)
\end{aligned}$$

where $\boldsymbol{\pi}_f$ is the initial MMSE estimate of the latent variable, $\boldsymbol{z}_f$ is the initial MSE, $\boldsymbol{q}_{f,\tau_1|\tau_2}$ is the MMSE estimate of $\boldsymbol{q}_{f,\tau_1}$ under the condition that $\boldsymbol{x}_{f,\tau}$ ($\tau = 1 \dots \tau_2$) are observed, and $\boldsymbol{R}_{f,\tau}$ is the MSE matrix.

Kalman smoother:

$$\begin{aligned}
\boldsymbol{R}_{f,\tau|L_T} &= \boldsymbol{R}_{f,\tau|\tau}\\
&\quad - \boldsymbol{B}_{f,\tau}(\boldsymbol{R}_{f,\tau+1|\tau} - \boldsymbol{R}_{f,\tau+1|L_T})\boldsymbol{B}^H_{f,\tau}, && (24)\\
\boldsymbol{q}_{f,\tau|L_T} &= \boldsymbol{q}_{f,\tau|\tau} + \boldsymbol{B}_{f,\tau}(\boldsymbol{q}_{f,\tau+1|L_T} - \boldsymbol{q}_{f,\tau+1|\tau}), && (25)\\
\boldsymbol{P}_{f,\tau} &= \boldsymbol{R}_{f,\tau|L_T} + \boldsymbol{q}_{f,\tau|L_T}\boldsymbol{q}^H_{f,\tau|L_T}, && (26)\\
\boldsymbol{P}_{f,\tau,\tau-1} &= \boldsymbol{R}_{f,\tau,\tau-1|L_T} + \boldsymbol{q}_{f,\tau|L_T}\boldsymbol{q}^H_{f,\tau-1|L_T}, && (27)\\
\boldsymbol{R}_{f,\tau,\tau-1|L_T} &= \boldsymbol{R}_{f,\tau|\tau}\boldsymbol{B}^H_{f,\tau-1}\\
&\quad - \boldsymbol{B}_{f,\tau}(\boldsymbol{A}_f\boldsymbol{R}_{f,\tau|\tau} - \boldsymbol{R}_{f,\tau+1,\tau|L_T})\boldsymbol{B}^H_{f,\tau-1}, && (28)\\
\boldsymbol{R}_{f,L_T,L_T-1|L_T} &= (\boldsymbol{I} - \boldsymbol{K}_{f,L_T}\boldsymbol{F})\boldsymbol{A}_f\boldsymbol{R}_{f,L_T-1|L_T-1}, && (29)\\
\boldsymbol{B}_{f,\tau} &= \boldsymbol{R}_{f,\tau|\tau}\boldsymbol{A}^H_f\boldsymbol{R}^{-1}_{f,\tau+1|\tau}, && (30)
\end{aligned}$$

The dereverberated noiseless speech signal is estimated as the following state-transition noise:

$$\hat{\boldsymbol{u}}_{f,\tau} = \boldsymbol{q}_{f,\tau|L_T} - \boldsymbol{A}_f\boldsymbol{q}_{f,\tau-1|L_T}, \tag{31}$$

where the first N_mth row of $\hat{\boldsymbol{u}}_{f,\tau}$ is the dereverberated speech source estimate.

3.3. Parameter optimization

In the M step, the parameters are updated by using the sufficient statistics obtained using the Kalman smoother. The optimization algorithm for the state-space model with the time-varying state-transition covariance matrix is slightly modified by the parameter optimization algorithm for the state-space model with the time-invariant state-transition covariance matrix cited by [8]. The linear prediction coefficients are obtained as

$$[\ \boldsymbol{W}_{f,D} \quad \dots \quad \boldsymbol{W}_{f,L_2-1}\ ] = \Big(\sum_{\tau=2}^{L_T} \frac{\boldsymbol{P}^{(2)}_{f,\tau,\tau-1}}{v_{f,\tau}}\Big)\Big(\sum_{\tau=2}^{L_T} \frac{\boldsymbol{P}^{(2)}_{f,\tau-1}}{v_{f,\tau}}\Big)^{-1}, \tag{32}$$

where $\boldsymbol{P}^{(2)}_{f,\tau-1}$ is the submatrix of $\boldsymbol{P}_{f,\tau-1}$ from the $\{N_m(D-1)+1\}$th column to the last column and from the $N_m(D-1)+1$th row to the last row, and $\boldsymbol{P}^{(2)}_{f,\tau,\tau-1}$ is the submatrix of $\boldsymbol{P}_{f,\tau,\tau-1}$ from the $\{N_m(D-1)+1\}$th column to the last column and from the first row to the N_mth row. The optimized linear prediction coefficients and sufficient statistics are used to obtain the remaining parameters:

$$v_t = \frac{1}{N_m}\mathrm{trace}\{\boldsymbol{R}^{-1}_u \boldsymbol{Q}^{(2)}_{f,\tau}\}, \tag{33}$$

$$\boldsymbol{R}_{u,f} = \frac{1}{L_T - 1}\sum_{t=2}^{L_T}\frac{1}{v_t}\boldsymbol{Q}^{(2)}_{f,\tau}, \tag{34}$$

$$\boldsymbol{R}_{w,f} = \frac{1}{L_T}\sum_{\tau=1}^{L_T}\{\boldsymbol{x}_{f,\tau}\boldsymbol{x}^H_{f,\tau} - \boldsymbol{F}\boldsymbol{q}_{f,\tau|L_T}\boldsymbol{x}^H_{f,\tau}\}, \tag{35}$$

where $\boldsymbol{Q}^{(2)}_{f,\tau}$ is the submatrix of $\boldsymbol{Q}_{f,\tau}$ from the first column to the N_mth column and from the first row to the N_mth row. $\boldsymbol{Q}_{f,\tau}$ is defined as

$$\boldsymbol{Q}_{f,\tau} = \boldsymbol{P}_{f,\tau} - \boldsymbol{A}_f\boldsymbol{P}^H_{f,\tau,\tau-1} - \boldsymbol{P}_{f,\tau,\tau-1}\boldsymbol{A}^H_f + \boldsymbol{H}\boldsymbol{P}_{f,\tau-1}\boldsymbol{A}^H_f. \tag{36}$$

The initial MMSE estimate and the initial MSE of the state vector are updated as follows:

$$\boldsymbol{\pi}_f = \boldsymbol{q}_{f,1|L_T}, \tag{37}$$

$$\boldsymbol{z}_f = \boldsymbol{R}_{f,1|L_T}. \tag{38}$$

4. EVALUATION

4.1. Setup

The proposed method was evaluated by using impulse responses measured in a room (Fig. 2, RT_{60} = 430 ms) and a microphone array with three elements. The impulse responses were measured at two locations, with the direction of sound radiation from the loudspeaker θ set to 0 degrees. The background noise was also recorded in the same environment using the same microphone array. Original source signals were extracted from the TIMIT database [11] for 34 speakers (one utterance each). The evaluation metric was SDR increase:

$$\begin{aligned}
\text{SDR increase} &= 10\log_{10}\frac{\sum_{n=1}^{L_T}\sum_{m=1}^{N_m} s_m(n)^2}{\sum_{n=1}^{L_T}\sum_{m=1}^{N_m}(s_m(n) - y_m(n))^2}\\
&\quad - 10\log_{10}\frac{\sum_{n=1}^{L_T}\sum_{m=1}^{N_m} s_m(n)^2}{\sum_{n=1}^{L_T}\sum_{m=1}^{N_m}(s_m(n) - x_m(n))^2},
\end{aligned}$$

where $x_m(n)$ is the mth microphone input signal at the nth point, $s_m(n)$ is the desired signal in the mth microphone at the nth point

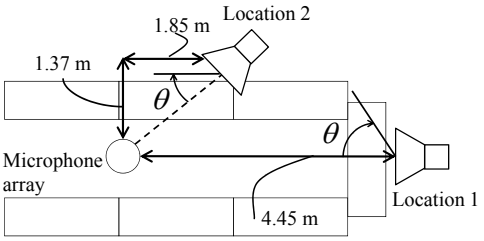

Fig. 2. Experimental environment: sound radiation direction θ set to 0 degrees.

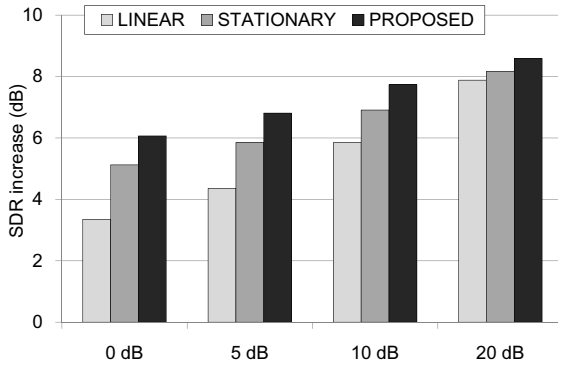

Fig. 3. Increase in SDR at location 1

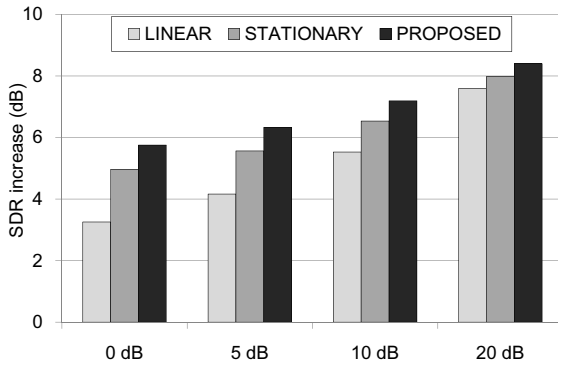

Fig. 4. Increase in SDR at location 2

(which is the summation of the direct sound and early reflection), and $y_m(n)$ is the dereverberated signal for the mth channel. The average value of ΔSNR for 34 utterances is used as the metric. The SNR of the microphone input signal was set to 0, 5, 10, or 20 dB. The other parameters were set as are shown in Table 1.

Table 1. Evaluation conditions

Sampling rate (Hz)	16,000
Frame size (pt)	1024
Frame shift (pt)	256
Number of microphones N_m	3
Number of EM iterations	20

The proposed method (PROPOSED) was compared with two other methods

- LINEAR: Dereverberation using linear prediction of microphone input signal. Covariance matrix of microphone input signal is modeled as $\boldsymbol{R}_{x,f,\tau} = v_{f,\tau}\boldsymbol{R}_{u,f}$, where $v_{f,\tau}$ and $\boldsymbol{R}_{u,f}$ are updated using EM algorithm.
- STATIONARY: Dereverberation using linear prediction of microphone input signal and multichannel Wiener-filtering-based stationary background noise reduction. Covariance matrix of microphone input signal is modeled as $\boldsymbol{R}_{x,f,\tau} = v_{f,\tau}\boldsymbol{R}_{u,f} + \boldsymbol{R}_{w,f}$ where $v_{f,\tau}$, $\boldsymbol{R}_{u,f}$, and $\boldsymbol{R}_{w,f}$ are updated using EM algorithm.

As shown in Figs. 3 and 4, the lower the SNR of the microphone input signal, the more effective the proposed method at each location of the speech source. This means the proposed method can reduce reverberation and the background noise effectively even when the microphone input signal is recorded in noisy environments.

5. CONCLUSION

Our proposed noise robust speech dereverberation method is based on linear prediction of the noiseless speech signal, which is estimated using the Kalman smoother with the time-varying covariance matrix as the state-transition covariance matrix. Noiseless speech signal estimation and parameter optimization are performed in an iterative manner on the basis of the EM algorithm. Testing showed that the proposed method is better than conventional methods under noisy conditions.

6. RELATION TO PRIOR WORK

Previous linear prediction based dereverberation methods [3][4][5] use a blind speech dereverberation technique: linear prediction of the microphone input signal. In contrast, our proposed method uses linear prediction of the noiseless speech signal to reduce degradation of speech dereverberation performance in noisy environments. While a time-varying covariance-matrix-based source separation method has been proposed [10], it is only for source separation without dereverberation. The optimization scheme used in the proposed method is based on optimization using the conventional Kalman smoother [8]. The conventional method uses a time-invariant covariance matrix for the state-transition covariance matrix while our proposed method uses a time-varying covariance matrix so as to meet the time-varying speech characteristics.

7. REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, vol. 30, no. 2, pp. 145-152, Feb. 1988.

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

[3] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of Late Reverberation Effect on Speech Signal Using Long-term Multiple-step Linear Prediction," *IEEE Trans. ASLP,* vol. 17, no. 4, pp. 534–545, 2009.

[4] T. Yoshioka, T. Nakatani, M. Miyoshi, and H.G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. ASLP*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[5] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Multichannel speech dereverberation and separation with optimized combination of linear and non-linear filtering," in *Proc. ICASSP2012, 2012*, pp. 4057–4060.

[6] A.P. Dempster et al., "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistic Society*, Series B 39(1),pp. 1–38, 1977.

[7] A.H. Jazwinski, *Stochastic Processes and Filtering Theory.* Academic Press, 1970.

[8] Z. Ghahramani and G.E. Hinton, "Parameter estimation for linear dynamical systems," Technical Report CRG-TR-96-2, Department of Computer Science, University of Toronto, 1996.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Speech Audio Process.*, vol. 17, no. 7, pp. 1717–1731, Sep. 2010.

[10] N.Q.K. Duong, E. Vincent, R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Speech Audio Process.*, vol. 18, no. 7, pp. 1830–1840, 2010/9.

[11] TIMIT corpus, http://www.ldc.upenn.edu/ Catalog/CatalogEntry.jsp?catalogId=LDC93S1.