FREQUENCY DOMAIN ACOUSTIC ECHO REDUCTION BASED ON KALMAN SMOOTHER WITH TIME-VARYING NOISE COVARIANCE MATRIX

Masahito Togami, Yohei Kawaguchi, and Ryoichi Takashima

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

ABSTRACT

In this paper, we propose a novel acoustic-echo-reduction technique at a time-frequency domain, which is optimally combined with speech enhancement. Unlike conventional echo reduction techniques which minimizes only residual power of the far-end acoustic echo signal, the proposed method minimizes summation of the residual echo signal and distortion of the near-end speech signal from a minimum mean square error (MMSE) perspective. The proposed method performs echo reduction with speech enhancement and parameter optimization in an iterative manner based on the expectation-maximization (EM) algorithm. The E step is corresponding with the echo reduction and speech enhancement based on the Kalman smoother with a time-varying covariance matrix for the observation noise term, which reflects the time-varying characteristics of speech sources. By using the time-varying covariance matrix, we can enhance speech sources effectively with acoustic echo reduction. Associated with the time-varying covariance matrix, a new optimization scheme of parameters for the M step is derived in this paper. Experimental results with impulse responses which was recorded under a real meeting room show that the proposed method can effectively enhance a near-end speech signal when there are a near-end speech signal and a far-end acoustic echo signal.

Index Terms— Time-varying assumption, acoustic echo reduction, EM algorithm

1. INTRODUCTION

Acoustic echo reduction techniques have been studied for a long time so as to avoid acoustic feedback in remote conference systems. Conventionally, there are many acoustic echo reduction techniques based on least square algorithms such as Normalized Least Mean Squares (NLMS) [1]. These techniques utilize a finite impulse response (FIR) filter for echo reduction which is optimized so as to minimize the averaged power of the residual signal after echo reduction. From a probabilistic perspective, minimization of the averaged power w.r.t the FIR filter is equivalent with maximum likelihood estimation of the FIR filter under the assumption that the probability density function (PDF) of the residual signal after filtering is a time-invariant Gaussian distribution. The time-invariant assumption for the PDF of the residual signal is reasonable when there is only background noise signal in the residual signal. However, in general there are both a near-end speech signal and the background noise signal in the residual signal, and the time-invariant assumption for the PDF of the near-end speech signal is not adequate, because speech signals are non-stationary signals. From a different perspective, the conventional least square algorithms focus on only echo reduction, and speech enhancement performance is under the sufficient level for remote conference systems.

In this paper, we propose a novel echo reduction technique which is an extension of a conventional single channel acoustic echo canceller with the Kalman filtering at the time-domain [2]. The proposed method extends the conventional method into a multichannel acoustic echo reduction technique at a time-frequency domain. By using multichannel microphone input signal, a multichannel beamforming technique can be integrated with acoustic echo reduction. In addition to the multichannel extension, the proposed method effectively integrate the time-varying assumption of speech sources with echo reduction by using a time-varying covariance matrix for the observation noise term in the observation equation. The time-varying covariance matrix model in the observation noise term is recently proposed local Gaussian modeling [3]. For parameter optimization, the proposed method utilizes a expectation-maximization (EM) based optimization scheme. The E step is corresponding with the Kalman smoother. We derive a novel parameter optimization scheme for the Kalman smoother with a time-varying covariance matrix for the observation noise term. By using the sufficient statistics that is estimated by the Kalman smoother, the parameters are updated so as to increase the Q function in the M step. In this paper, the proposed method is formulated as an offline echo reduction technique. However, by using an online EM method based parameter optimization scheme for local Gaussian modeling [4], the proposed method can be extended into an online echo reduction technique.

The proposed method is evaluated by two scenarios, a synchronization case of an A/D converter and a D/A converter and an asynchronous case. Evaluation results show that the proposed method can more effectively reduce acoustic echo than the conventional method. even when the A/D converter and the D/A converter are not synchronized with each other.

2. PROBLEM STATEMENT

2.1. Input signal model

In this paper, acoustic echo reduction is performed at a time-frequency domain by using shor-term Fourier transform. The microphone input signal at the time-frequency domain is expressed as follows:

$$\boldsymbol{x}(l,k) = \sum_{n=1}^{N_s} \boldsymbol{c}_n(l,k) + \sum_{l'=0}^{L_d-1} \boldsymbol{h}_{l'}(l,k) d(l-l',k) + \boldsymbol{w}(l,k), \quad (1)$$

where l is the frame index, k is the frequency index, N_s is the number of the sources, $\boldsymbol{c}_n(l,k)$ is the nth source signal, L_d is the tap-length of an acoustic impulse response at the time-frequency domain, $\boldsymbol{h}_{l'}(l,k)$ is the l'th tap of the acoustic impulse response at each time-frequency point, and $\boldsymbol{w}(l,k)$ is a multichannel background noise signal. $d(l,k)$ is the far-end speech signal at each

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

time-frequency point. The acoustic impulse response is modeled as a gradually time-varying variable as follows:

$$\boldsymbol{h}_{l'}(l,k) = \boldsymbol{h}_{l'}(l-1,k) + \boldsymbol{e}_{l'}(l,k), \tag{2}$$

where $\boldsymbol{e}_{l'}(l,k)$ is the amount of change of the acoustic impulse response. In the teleconferencing systems, $d(l,k)$ is known in advance, because the far-end speech signal is transferred from a remote site. In this paper, the goal of the echo reduction is set to extraction of $\sum_{n=1}^{N_s} \boldsymbol{c}_n(l,k)$ in (1) from the microphone input signal $\boldsymbol{x}(l,k)$ and the far-end speech signal $d(l,k)$.

2.2. State-space model

Similar to the conventional single channel echo reduction technique with a Kalman-filtering technique [2], we derive a Kalman-smoother based acoustic echo reduction technique by modifying the original microphone input model (1) into a state-space model and an observation model. On contrary to the conventional single channel method, the proposed method utilizes a multichannel input model. Multichannel microphone input signal model defined by (1) can be modified into a following equation:

$$\boldsymbol{x}(l,k) = \boldsymbol{D}(l,k)\boldsymbol{H}(l,k) + \boldsymbol{G}(l,k), \tag{3}$$

where

$$\boldsymbol{D}(l,k) = \Big[\ d(l,k)\boldsymbol{I}_{N_m\times N_m} \quad d(l-1,k)\boldsymbol{I}_{N_m\times N_m} \quad \dots \quad d(l-L_d+1,k)\boldsymbol{I}_{N_m\times N_m}\ \Big]. \tag{4}$$

$\boldsymbol{I}_{N_m\times N_m}$ is set to a $N_m \times N_m$ identity matrix. The impulse response is summarized as follows:

$$\boldsymbol{H}(l,k) = [\ \boldsymbol{h}_0(l,k)^T \quad \boldsymbol{h}_1(l,k)^T \quad \dots \quad \boldsymbol{h}_{L_d-1}(l,k)^T\ ]^T, \tag{5}$$

where T is a transpose operator of a matrix/vector. $\boldsymbol{G}(l,k)$ is set to summation of a near-end speech signal and a background noise signal as follows:

$$\boldsymbol{G}(l,k) = \sum_{n=1}^{N_s} \boldsymbol{c}_n(l,k) + \boldsymbol{w}(l,k). \tag{6}$$

(3) can be regarded as an observation equation of Kalman filter [5]. $\boldsymbol{H}(l,k)$ can be expressed as a following state-transition equation:

$$\boldsymbol{H}(l,k) = \boldsymbol{H}(l-1,k) + \boldsymbol{E}(l,k), \tag{7}$$

where

$$\boldsymbol{E}(l,k) = [\ \boldsymbol{e}_0(l,k)^T \quad \boldsymbol{e}_1(l,k)^T \quad \dots \quad \boldsymbol{e}_{L_d-1}(l,k)^T\ ]^T. \tag{8}$$

3. PROPOSED METHOD

3.1. Probabilistic modeling

In the proposed method, all of the variables in the state-transition equation and the observation equation are modeled as time-varying or time-invariant Gaussian distributions.

3.1.1. Probabilistic model for near-end speech signals

Similar to [3], the PDF of the near-end speech signal $\boldsymbol{c}_n(l,k)$ is modeled as a time-varying Gaussian distribution as follows:

$$p(\boldsymbol{c}_n(l,k)) = \mathcal{N}(\boldsymbol{c}_n(l,k); \boldsymbol{0}, v_n(l,k)\boldsymbol{R}_n(k)), \tag{9}$$

where $v_n(l,k)$ is a time-varying scalar coefficient of the nth speech source, $\boldsymbol{R}_n(k)$ is the time-invariant covariance matrix of the acoustic transfer function of the nth speech source.

3.1.2. Probabilistic model for background noise signal

The background noise signal is modeled as a stationary Gaussian distribution as follows:

$$\boldsymbol{w}(l,k) = \mathcal{N}(\boldsymbol{w}(l,k); \boldsymbol{0}, \boldsymbol{R}_w(k)), \tag{10}$$

where $\boldsymbol{R}_w(k)$ is modeled as a time-invariant covariance matrix of the background noise signal.

3.1.3. Probabilistic model for state-transition noise

The state-transition noise term $\boldsymbol{E}(l,k)$ is modeled as the following time-invariant Gaussian distribution:

$$p(\boldsymbol{E}(l,k)) = \mathcal{N}(\boldsymbol{E}(l,k); \boldsymbol{0}, \sigma(k)\boldsymbol{I}_{N_m\times N_m}), \tag{11}$$

where $\sigma(k)$ is the amount of change of the acoustic impulse response at each time-frequency point.

3.2. Sufficient statistics estimation

Acoustic echo reduction with speech enhancement and parameter optimization are performed in an iterative manner based on the EM algorithm [6]. The E step is corresponding with sufficient statistics of probabilistic variables. Under the derived state-space model and the derived observation model, at first, the proposed method estimates the sufficient statistics of the probabilistic variables. By using the estimated sufficient statistics, the proposed method updates the parameters so as to increase the likelihood function.

3.2.1. Sufficient statistics for acoustic impulse response

The Kalman smoother is utilized for estimation of the latent variables $\boldsymbol{H}(l,k)$. The sufficient statistics are

$$\begin{aligned}
\boldsymbol{H}_{l|L_T,k} &= \mathrm{E}[\boldsymbol{H}(l,k)|\mathcal{X}(k),\mathcal{D}(k)], && (12)\\
\boldsymbol{R}_{l|L_T,k} &= \mathrm{E}[(\boldsymbol{H}(l,k)-\boldsymbol{H}_{l|L_T,k}) \\
&\times (\boldsymbol{H}(l,k)-\boldsymbol{H}_{l|L_T,k})^H|\mathcal{X}(k),\mathcal{D}(k)], && (13)\\
\boldsymbol{R}_{l,l-1|L_T,k} &= \mathrm{E}[(\boldsymbol{H}(l,k)-\boldsymbol{H}_{l|L_T,k}) \\
&\times (\boldsymbol{H}(l-1,k)-\boldsymbol{H}_{l-1|L_T,k})^H|\mathcal{X}(k),\mathcal{D}(k)], && (14)
\end{aligned}$$

where H is the Hermite transpose of a matrix/vector, E is the operator for the expected value calculation, $\cdot_{l|L_T,k} = E[\cdot(l,k)|\mathcal{X}(k),\mathcal{D}(k)]$, L_T is the number of the frames, $\mathcal{X}(k) = \{\boldsymbol{x}(1,k),\dots,\boldsymbol{x}(L_T,k)\}$, and $\mathcal{D}(k) = \{d(1,k),\dots,d(L_T,k)\}$. At first, the Kalman filtering [5] is utilized as follows:

Prediction

$$\boldsymbol{H}_{l|l-1,k} = \boldsymbol{H}_{l-1|l-1,k} \tag{15}$$

Kalman filtering

$$\boldsymbol{H}_{l|l,k} = \boldsymbol{H}_{l|l-1,k} + \boldsymbol{K}(l,k)\Big(\boldsymbol{x}(l,k) - \boldsymbol{D}(l,k)\boldsymbol{H}_{l|l-1,k}\Big), \tag{16}$$

where $\boldsymbol{K}(l,k)$ is a Kalman gain, which is calculated as follows:

$$\boldsymbol{K}(l,k) = \boldsymbol{R}_{l|l-1,k}\boldsymbol{D}(l,k)^H\boldsymbol{R}_x(l,k)^{-1}, \tag{17}$$

$$\boldsymbol{R}_{l|l-1,k} = \boldsymbol{R}_{l-1|l-1,k} + \boldsymbol{R}_v(k) \tag{18}$$

$$\boldsymbol{R}_x(l,k) = \boldsymbol{R}_W(l,k) + \boldsymbol{D}(l,k)\boldsymbol{R}_{l|l-1,k}\boldsymbol{D}(l,k)^H \tag{19}$$

$\boldsymbol{R}_W(l,k)$ is summation of a covariance matrix of a near-end speaker and a covariance matrix of a background noise, which calculated as follows:

$$\boldsymbol{R}_W(l,k) = \sum_{n=1}^{N_s} v_n(l,k)\boldsymbol{R}_n(k) + \boldsymbol{R}_w(k) \tag{20}$$

Expected value of square error of $\boldsymbol{H}$ at the lth frame is calculated as follows:

$$\boldsymbol{R}_{l|l,k} = \Big(\boldsymbol{I} - \boldsymbol{K}(l,k)\boldsymbol{D}(l,k)\Big)\boldsymbol{R}_{l|l-1,k}. \tag{21}$$

Next, the MMSE estimate at the lth frame and the MSE is calculated by using Kalman smoother [7].

3.2.2. *Sufficient statistics for parameters of near-end speech source and background noise*

The sufficient statistics for the near-end speech source $\boldsymbol{c}_n(l,k)$ and the background noise signal $\boldsymbol{w}(l,k)$ are the minimum mean square error (MMSE) estimates and the mean square error (MSE) matrices. These statistics are estimated by using $\boldsymbol{H}_{l|L_T,k}$ and $\boldsymbol{R}_{l|L_T,k}$ which is estimated by the Kalman smoother as follows.

$$\begin{aligned}\hat{\boldsymbol{c}}_{n,l|L_T,k} &= \int \boldsymbol{c}_n(l,k)p(\boldsymbol{c}_n(l,k)|\mathcal{X}(k),\mathcal{D}(k))d\boldsymbol{c}_n(l,k) &(22)\\ &= \boldsymbol{A}_{n,l,k}\Big(\boldsymbol{x}(l,k) - \boldsymbol{D}(l,k)\boldsymbol{H}_{l|L_T,k}\Big), &(23)\end{aligned}$$

where $\boldsymbol{A}_{n,l,k} = v_n(l,k)\boldsymbol{R}_n(k)\boldsymbol{R}_W(l,k)^{-1}$ is a multichannel Wiener filter which extracts $\boldsymbol{c}_n(l,k)$ in the microphone input signal $\boldsymbol{x}(l,k)$. The MSE matrix of the nth near-end speech signal is obtained as follows:

$$\begin{aligned}\boldsymbol{R}_{c_n,l|L_T,k} &= \int \boldsymbol{c}_n(l,k)\boldsymbol{c}_n(l,k)^H\\ &\quad\times p(\boldsymbol{c}_n(l,k)|\mathcal{X}(k),\mathcal{D}(k))d\boldsymbol{c}_n(l,k)\\ &\quad-\hat{\boldsymbol{c}}_{n,l|L_T,k}\hat{\boldsymbol{c}}_{n,l|L_T,k}^H\\ &= \boldsymbol{A}_{n,l,k}\boldsymbol{Q}_{l,k|1\ldots L_T}\boldsymbol{A}_{n,l,k}^H\\ &\quad+(\boldsymbol{I} - \boldsymbol{A}_{n,l,k})v_n(l,k)\boldsymbol{R}_n(k), &(24)\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{Q}_{l,k|1\ldots L_T} &= \boldsymbol{x}(l,k)\boldsymbol{x}(l,k)^H - \boldsymbol{D}(l,k)\boldsymbol{H}_{l|L_T}\boldsymbol{x}(l,k)^H\\ &\quad-\boldsymbol{x}(l,k)\boldsymbol{H}_{l|L_T}^H\boldsymbol{D}(l,k)^H + \boldsymbol{D}(l,k)\boldsymbol{P}_{l,k}\boldsymbol{D}(l,k)^H, &(25)\end{aligned}$$

$$\boldsymbol{P}_{l,k} = \boldsymbol{R}_{l|L_T,k} + \boldsymbol{H}_{l|L_T,k}\boldsymbol{H}_{l|L_T,k}^H. \tag{26}$$

Similar to the near-end speech source case, the MMSE estimate and the MSE matrix of the background noise signal are estimated as follows:

$$\hat{\boldsymbol{w}}_{l|L_T,k} = \boldsymbol{B}_{l,k}\Big(\boldsymbol{x}(l,k) - \boldsymbol{D}(l,k)\boldsymbol{H}_{l|L_T,k}\Big), \tag{27}$$

$$\boldsymbol{R}_{w,l|L_T,k} = \boldsymbol{B}_{l,k}\boldsymbol{Q}_{l,k|1\ldots L_T}\boldsymbol{B}_{l,k}^H + (\boldsymbol{I} - \boldsymbol{B}_{l,k})\boldsymbol{R}_w(k). \tag{28}$$

where $\boldsymbol{B}_{l,k} = \boldsymbol{R}_w(k)\boldsymbol{R}_W(l,k)^{-1}$ is a multichannel Wiener filter which extracts $\boldsymbol{w}(l,k)$ in the microphone input signal $\boldsymbol{x}(l,k)$.

3.3. Parameter optimization (M step)

In the M step, the proposed method estimates the parameters so as to increase the Q function. By using estimated sufficient statistics in the E step, the parameters are updated so as to increase the Q function as follows:

$$v_n(l,k) = \frac{1}{N_m}\text{trace}\Big\{\boldsymbol{R}_n(k)^{-1}\boldsymbol{R}_{c_n,l|L_T,k}\Big\}, \tag{29}$$

$$\boldsymbol{R}_n(k) = \frac{1}{L_T}\sum_{l=1}^{L_T}\frac{\boldsymbol{R}_{c_n,l|L_T,k}}{v_n(l,k)}, \tag{30}$$

$$\boldsymbol{R}_w(k) = \frac{1}{L_T}\sum_{l=1}^{L_T}\boldsymbol{R}_{w,l|L_T,k}, \tag{31}$$

$$\sigma(k) = \frac{1}{N_m(L_T-1)}\text{tr}\left\{\sum_{l=2}^{L_T}\Big(\boldsymbol{P}_{l,k} - \boldsymbol{P}_{l,l-1,k}^H - \boldsymbol{P}_{l,l-1,k} + \boldsymbol{P}_{l-1,k}\Big)\right\}, \tag{32}$$

where $\boldsymbol{P}_{l,l-1,k} = \boldsymbol{R}_{l,l-1|L_T,k} + \boldsymbol{H}_{l|L_T,k}\boldsymbol{H}_{l-1|L_T,k}^H$ and tr is the operator that returns the trace value of a matrix/vector.

4. EXPERIMENT

4.1. Experimental condition

The experimental environment and the microphone array alignment are shown in Fig. 1. The impulse responses for a near-end speech signal were measured at the Location 1, 2, and 3 by using the TSP (Time-Stretched Pulse) method [8]. The near-end speech is generated by convoluting the measured impulse responses with the original source signal. The far-end speech signal and background noise were recording by using the same microphone array that was utilized for measuring the impulse responses of the near-end speech signal. At the first experiment, we evaluated echo reduction and speech enhancement performance when an A/D converter and a D/A converter are synchronized (Time-invariant echo-path case). Generally speaking, in the teleconferencing scene in large rooms, an A/D converter

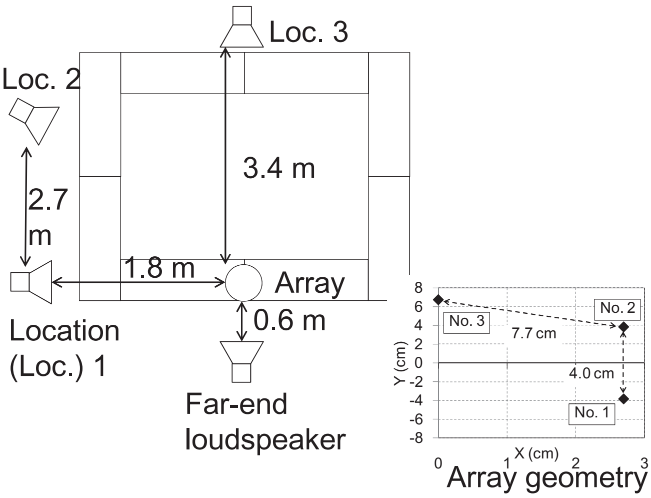

Fig. 1. Experimental environment and microphone array alignment

and a D/A converter are not always synchronized. In the second experiment, we evaluate the acoustic echo reduction performance when an A/D converter and a D/A converter are not synchronized. The far-end speech is played by using the D/A converter attached with the personal computer. Recording of the microphone input signal is performed by using the A/D converter which is not synchronized with the D/A converter (Time-varying echo-path case). The original source signals of the near-end speech signals and the far-end speech signals are extracted from the TIMIT database [9]. The number of the near-end speech signals and the number of the far-end speech signals are 34 each. The other experimental conditions are shown in Table. 1.

Table 1. Experimental conditions

Sampling rate	16000 [Hz]
Frame size	1024 [pt]
Frame shift	256 [pt]
Number of microphones N_m	3
L_d	8 [tap]
Number of EM iterations	10
N_s	1

Signal to Noise ratio (SNR) between the near-end speech and summation of the recorded far-end speech and the background noise signal was set to 0 dB. The evaluation measure is PESQ [11].

The following 3 methods based on the proposed method are comparatively evaluated.

1. INVARIANT: The time-invariant value $v_n(l, k)$ is set to 1. The background noise reduction is not performed, $\boldsymbol{R}_w(l, k) = \boldsymbol{0}$.
2. VARIANT: The variance of the near-end speech is set to a time-varying value. The background noise reduction is not performed, $\boldsymbol{R}_w(l, k) = \boldsymbol{0}$.
3. VARIANT+NC:The variance of the near-end speech is set to a time-varying value. The background noise reduction is performed.

4.2. Experimental results

4.2.1. Time-invariant echo-path case

Experimental results when the A/D converter and the D/A converter are synchronized are shown in Fig. 2 for each location. VARIANT+NC achieved the best performance at each location. VARIANT+NC and VARIANT achieved higher PESQ than INVARIANT, which means that the proposed time-varying model for the observation noise term is effective. Evaluation results for PESQ are shown in Fig. 2.

4.2.2. Time-varying echo-path case

Experimental results for PESQ when the A/D converter and the D/A converter are not synchronized are shown in Fig. 3 for each location. VARIANT+NC achieved the best performance. VARIANT+NC higher PESQ than VARIANT by 0.06 pt. The proposed method is shown to be effective even when the A/D converter and the D/A converter are not synchronized. In addition to tracking to the echo-path change, the proposed method can separate the far-end speech signal and the near-end speech signal spatially with multiple microphones.

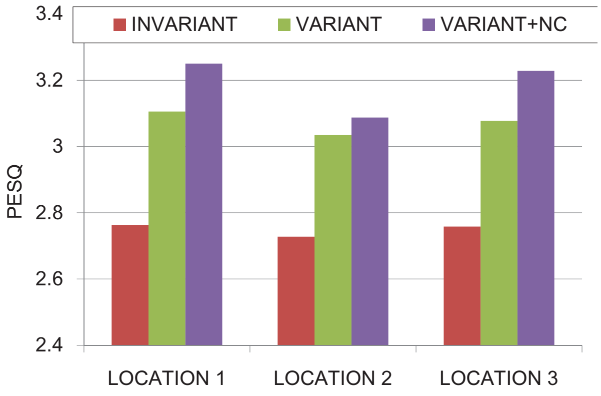

Fig. 2. Evaluation results of PESQ when A/D converter and D/A converter are synchronized

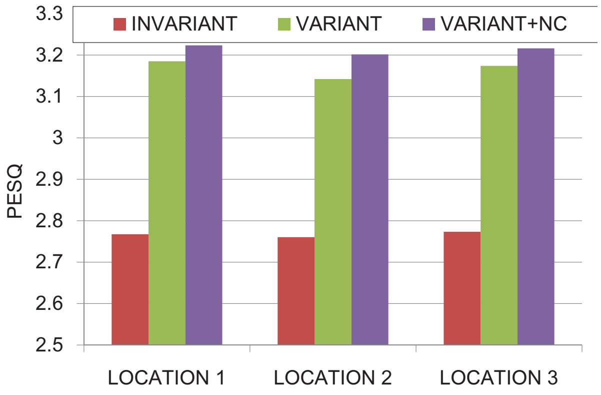

Fig. 3. Experimental results for PESQ: SNR is set to 0 dB.

5. CONCLUSION

We proposed a frequency domain echo reduction technique which is based on the Kalman smoother with time-varying characteristics of the near-end speech signal. The experimental results show that the proposed method can reduce acoustic echo and background noise signal effectively under noisy environments.

6. RELATION TO PRIOR WORK

Conventionally, authors proposed a probabilistic optimization approach of an echo reduction technique and a source separation technique [12] with time-varying assumption of speech sources. However, this approach assume that PDF of the far-end echo path is a stationary Gaussian distribution. When the far-end echo path gradually changes, the PDF of the far-end echo path also gradually changes, e.g., asynchronous cases for a A/D converter and a D/A converter case. In this case, echo reduction technique of the conventional method degrades. However, the proposed method tracks more accurately the change of the echo-path under the assumption that the echo-path gradually changes.

7. REFERENCES

[1] E. Hänsler, G. Schmidt, “Acoustic Echo and Noise Control: A Practical Approach,”, John Wiley & Sons, 2004.

[2] C. Paleologu, J. Benesty, S. Ciochina, “Study of the General Kalman Filter for Echo Cancellation,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 8, pp. 1539–1549, 2013/8.

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

[4] M. Togami, “Online speech source separation based on maximum likelihood of local Gaussian modeling,” *Proc. IEEE ICASSP2011*, pp. 213–216, 2011.

[5] R.E. Kalman, “A new approach to linear filtering and prediction problems, ” *Trans. ASME, J. Basic Eng.*, vol. 82 D, no. 1, pp. 34–45, 1960.

[6] A.P. Dempster, N. Laird, and D.B. Rubin, “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] Y. Suzuki, F. Asano, H.Y. Kim, and T. Sone, “An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses,” *J. Acoust. Soc. Amer.* vol. 97, no. 2, pp. 1119–1123, Feb. 1995.

[9] TIMIT corpus [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1.

[10] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, “Optimized Speech Dereverberation From Probabilistic Perspective for Time Varying Acoustic Transfer Function,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 7, pp. 1369–1380, 2013/7.

[11] ITU-T P.862, “Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders,” ITU-T, 2001.

[12] M. Togami and K. Hori, “Multichannel Semi-Blind Source Separation Via Local Gaussian Modeling for Acoustic Echo Reduction,” in *Proc. Eur. Signal Process. Conf.*, 2011, pp. 496–500.