MULTICHANNEL SEMI-BLIND SOURCE SEPARATION VIA LOCAL GAUSSIAN MODELING FOR ACOUSTIC ECHO REDUCTION

Masahito Togami † ‡ *and Koichi Hori* ‡

† Central Research Laboratory, Hitachi Ltd.
‡ The University of Tokyo
† 1-280 Higashi-koigakubo, kokubunji-shi, Tokyo-to 185-8601, Japan
phone: + (81) 42-323-1111, fax: + (81) 42-327-7823, email: masahito.togami.fe@hitachi.com

ABSTRACT

In this paper, we propose a semi-blind source separation method under the assumption that the original signal of a certain source is given in advance. Main purpose of the proposed method is acoustic echo reduction. The proposed method can be regarded as an optimal coupling of linear echo canceller, multichannel source separation, and nonlinear echo suppressor under a newly proposed semi-blind local Gaussian modeling. We analyze the behavior of the proposed method when phase component of the acoustic transfer function of a loudspeaker is fluctuated. It is shown that the proposed method is more robust against phase fluctuation than the conventional linear echo canceller. Furthermore, the proposed method is superior to full-rank local Gaussian modeling and a cascade method of the linear echo canceller and the full-rank local Gaussian modeling.

1. INTRODUCTION

Acoustic echo reduction technique [1] is one of essential functions in TV conferencing systems. Many echo reduction techniques have been studied. Conventional approaches can be divided into linear echo cancellers [1] and nonlinear echo suppressors. When acoustic transfer function between a loudspeaker and a microphone is linear, theoretically, linear echo cancellers are the best solution for acoustic echo reduction. However, linear echo cancellers are sensitive to change of the acoustic transfer function. When the acoustic transfer function fluctuates in a short period, the linear echo cancellers can not reduce acoustic echo sufficiently. Especially, the linear echo cancellers are fragile to phase fluctuation of the acoustic transfer function. If distance between the loudspeaker and the microphone move by about 4 cm, phase component of 4000Hz will be reversed. In the reverberant room, phase fluctuation frequently occurs due to reverberation. In this case, the nonlinear echo suppressors based on single channel noise reduction techniques [2] such as Wiener filtering and spectral subtraction are preferable [3]. An alternative way for residual noise reduction is an integrated method of the linear acoustic echo canceller and multichannel source separation such as a generalized sidelobe canceller [4] or semi-blind source separation [5]. These techniques optimize linear echo canceller and multichannel-beamforming simultaneously. However, these techniques remain linear filtering techniques. One of authors proposed multichannel acoustic echo reduction technique by using amplitude difference between microphones [10]. However, this technique is only a nonlinear filtering.

In this paper, we propose an optimal coupling of linear echo canceller, multichannel spatial beamforming, and single channel echo suppressor. The proposed method is an extension of local Gaussian modeling (LGM) [6][7][8][9]. LGM approximates the multichannel covariance matrix of each source as multiplication of a time-variant scalar coefficient and a time-invariant matrix. The time-invariant matrix is related to direction of arrival of each source. Duong et al. proposed approximation of the time-invariant matrix as a full-rank matrix [7]. LGM with the full-rank matrices is shown to be robust against reverberation i.e. phase fluctuation. However, the original LGM cannot be applied for acoustic echo reduction problem. In this paper, LGM is extended in the following respect to apply for acoustic echo reduction problem:

1. The original LGM is blind source separation under the condition that original signals of all sources are unknown. We extend LGM for semi-blind source separation under the condition that a certain source is given in advance.
2. The original LGM neglects late reverberation and approximates microphone input signal as instantaneous mixtures at time-frequency domain. However, late reverberation of loudspeaker signal cannot be neglected, because it causes acoustic feedback. We extends LGM for convolutive mixtures of the loudspeaker signal in the time-frequency domain.

The original LGM regards probability distribution function of each source as a Gaussian distribution with 0-mean and full-rank covariance matrix. Each source is assumed to be independent of one another. In convolutive mixtures case, acoustic echo component of each tap is not independent of one another. The proposed method models probability distribution function of loudspeaker signal as a Gaussian distribution with non 0-mean and full-rank covariance matrix. Non 0-mean component is related to linear part of acoustic echo (dependent part of each tap), and full-rank covariance matrix is related to residual echo (independent part of each tap). Optimization scheme of the proposed method is based on the EM algorithm [11].

Experimental results show that the proposed method is robust against phase fluctuation, and the proposed method is superior to the linear echo canceller or the full-rank local Gaussian modeling.

2. PROBLEM STATEMENT

The m-th microphone input signal $x_m(t)$ is converted into time-frequency domain by using short term Fourier transform. In the microphone input signal at each time-frequency point is defined as follows:

$$\boldsymbol{x}_{f,\tau} = [\ x_{1,f,\tau} \quad \cdots \quad x_{M,f,\tau}\ ], \tag{1}$$

where f is the frequency index, τ is the frame index, and M is the number of the microphones. The microphone input signal is composed of near-end speech signal and far-end acoustic echo. Therefore, $\boldsymbol{x}_{f,\tau}$ can be modeled under the narrowband assumption as follows:

$$\boldsymbol{x}_{f,\tau} = \sum_{i=1}^{N_s} s_{i,f,\tau}\boldsymbol{a}_{i,f,\tau} + \sum_{j=1}^{L_d} d_{f,\tau-j+1}\boldsymbol{b}_{j,f,\tau} \tag{2}$$

where $\boldsymbol{a}_{i,f,\tau}$ is steering vector of the i-th near end speech, $s_{i,f,\tau}$ is the source signal of the i-th near-end speech, N_s is the number of near-end speech sources, $d_{f,\tau}$ is a reference signal of far-end speech, and $\boldsymbol{b}_{j,f,\tau}$ is the j-th transfer function of the far-end speech. For far-end speech signal, reverberation cannot be neglected, because reverberation causes acoustic feedback. Therefore, the far-end speech signal is modeled as a convolutive mixture in Eq. 2. In the echo cancellation problem, the reference signal $d_{f,\tau}$ can be assumed to be given. Purpose of the acoustic echo canceller is to remove acoustic echo component from microphone input signal $\boldsymbol{x}_{f,\tau}$ by using the given

$d_{f,\tau}$. All variables i.e. $s_{i,f,\tau}$, $\boldsymbol{a}_{i,f,\tau}$, $d_{f,\tau}$, $\boldsymbol{b}_{j,f,\tau}$ are assumed to be independent of each other. However, $d_{f,\tau}$, $d_{f,\tau-1},\ldots$ are correlated with each other because these variables are caused by the same source component. Furthermore, the source signals i.e. $s_{i,f,\tau}$ and $d_{f,\tau}$ are assumed to be 0-mean.

2.1 Local Gaussian Modeling

The microphone input signal can be converted as follows:

$$\boldsymbol{x}_{f,\tau} = \sum_{i=1}^{N_s} \boldsymbol{c}_{i,f,\tau} + \sum_{j=1}^{L_d} \boldsymbol{e}_{j,f,\tau}, \tag{3}$$

where

$$\begin{aligned} \boldsymbol{c}_{i,f,\tau} &= s_{i,f,\tau}\boldsymbol{a}_{i,f,\tau}, & (4)\\ \boldsymbol{e}_{j,f,\tau} &= d_{f,\tau-j+1}\boldsymbol{b}_{j,f,\tau}, & (5) \end{aligned}$$

$\boldsymbol{c}_{i,f,\tau}$ is the multichannel signal of the i-th source, $\boldsymbol{e}_{j,f,\tau}$ is the multichannel residual echo signal associated with $d_{f,\tau-j+1}$. Duong et al., [7] proposed a blind source separation technique which estimates the maximum likelihood value of $\boldsymbol{c}_{i,f,\tau}$ by the EM algorithm under the condition that $\boldsymbol{c}_{1,f,\tau}\ldots\boldsymbol{c}_{N_s,f,\tau}$ are mutually independent. Duong's method uses the approximation that the probability function of $\boldsymbol{c}_{i,f,\tau}$ is a Gaussian distribution with 0-mean and a multichannel covariance matrix which is multiplication of a time-variant scale factor and a time-invariant matrix (LGM: local Gaussian modeling).

2.2 Straightforward semi-blind LGM by using 0-mean and full-rank covariance model

The most straightforward approach for semi-blind source separation is a simple extension of LGM with 0-mean and full-rank covariance matrix under the assumption that $\boldsymbol{c}_{1,f,\tau}\ldots\boldsymbol{c}_{N_s,f,\tau}$, $\boldsymbol{e}_{j,f,\tau}\ldots\boldsymbol{e}_{j,f,\tau-L_d+1}$ are mutually independent and are Gaussian distributions with 0-mean and a multichannel covariance matrix. By following this approach, the multichannel covariance matrix of the microphone input signal can be approximated as sum of multichannel covariance matrix of each source. However, this approximation is problematic, because actually, $d_{f,\tau-j+1}$ is a deterministic variable, and expected value of cross spectral of $\boldsymbol{e}_{j,f,\tau}$ and $\boldsymbol{e}_{j_2,f,\tau}$, $\mathbb{E}[\boldsymbol{e}_{j,f,\tau}\boldsymbol{e}^H_{j_2,f,\tau}]$ is not 0 as follows:

$$\begin{aligned} & \mathbb{E}[\boldsymbol{e}_{j,f,\tau}\boldsymbol{e}^H_{j_2,f,\tau}] \\ = \; & \mathbb{E}[d_{f,\tau-j+1}d^*_{f,\tau-j_2+1}\boldsymbol{b}_{j,f,\tau}\boldsymbol{b}^H_{j_2,f,\tau}] \\ = \; & d_{f,\tau-j+1}d^*_{f,\tau-j_2+1}\hat{\boldsymbol{b}}_{j,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}\hat{\boldsymbol{b}}^H_{j_2,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} \neq 0, \end{aligned} \tag{6}$$

where $\mathbb{E}[\cdot]$ is an operator for mathematical expectation.

3. PROPOSED METHOD

3.1 Overview of proposed method

The local Gaussian modeling cannot be straightforwardly applied for acoustic echo reduction problem, because cross spectral of $\boldsymbol{e}_{1,f,\tau}\ldots\boldsymbol{e}_{L_d,f,\tau}$ are not zero. To overcome this situation, the proposed method uses decorrelation of $\boldsymbol{e}_{j,f,\tau}$ by using linear filtering as follows:

$$\hat{\boldsymbol{e}}_{j,f,\tau} = \boldsymbol{e}_{j,f,\tau} - d_{f,\tau-j+1}\mathbb{E}[\boldsymbol{b}_{j,f,\tau}]. \tag{7}$$

Cross spectral of $\hat{\boldsymbol{e}}_{1,f,\tau}\ldots\hat{\boldsymbol{e}}_{L_d,f,\tau}$ are expected to be zero. $\hat{\boldsymbol{e}}_{1,f,\tau}\ldots\hat{\boldsymbol{e}}_{L_d,f,\tau}$ are mutually independent and are Gaussian distributions with 0-mean and a multichannel covariance matrix. Therefore, if the microphone input signal is composed of $\boldsymbol{c}_{1,f,\tau}\ldots\boldsymbol{c}_{N_s,f,\tau}$, $\hat{\boldsymbol{e}}_{j,f,\tau}\ldots\hat{\boldsymbol{e}}_{j,f,\tau-L_d+1}$, the multichannel covariance matrix of the microphone input signal can be approximated as sum of covariance matrices of $\boldsymbol{c}_{1,f,\tau}\ldots\boldsymbol{c}_{N_s,f,\tau}$, $\hat{\boldsymbol{e}}_{j,f,\tau}\ldots\hat{\boldsymbol{e}}_{j,f,\tau-L_d+1}$. Therefore, distribution function of loudspeaker signal can be modeled as a Gaussian distribution with non 0-mean and full-rank covariance matrix.

In accordance with the above discussion, we propose an extended version of Duong' method to semi-blind source separation case with non-0 mean and full-rank covariance matrix of loudspeaker signal. Semi-blind source separation case with local Gaussian modeling is different from the original blind source separation technique, because the original blind source separation method only estimates the covariance matrix of each source, but the proposed method must estimate the covariance matrix and the average value of acoustic transfer function of the far-end source simultaneously.

3.2 Likelihood function

The log likelihood function which is intended to be maximized by the proposed method is defined as follows:

$$\begin{aligned} L(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{D}) = & \sum_f\sum_\tau\Big\{-(\boldsymbol{x}_{f,\tau}-\boldsymbol{\mu}_{f,\tau|\boldsymbol{\theta},\boldsymbol{D}})^H\boldsymbol{R}^{-1}_{x,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}(\boldsymbol{x}_{f,\tau}-\boldsymbol{\mu}_{f,\tau|\boldsymbol{\theta},\boldsymbol{D}}) \\ & - \log|\boldsymbol{R}_{x,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}| - \log|2\pi|^{\frac{M}{2}},\Big\} \end{aligned} \tag{8}$$

where $\boldsymbol{X} = \{\boldsymbol{x}_{f,\tau}\}$, $\boldsymbol{D} = \{\boldsymbol{d}_{f,\tau}\}$, H is an operator for Hermitian transpose of a vector/matrix, $\boldsymbol{\theta}$ is a model parameter vector which is required to be optimized. $\boldsymbol{\mu}_{f,\tau|\boldsymbol{\theta},\boldsymbol{D}}$ is the average vector of the microphone input signal at the frame τ and the frequency f, and $\boldsymbol{R}_{x,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}$ is the covariance matrix of the microphone input signal at the corresponding time-frequency point. $\boldsymbol{\mu}_{f,\tau|\boldsymbol{\theta},\boldsymbol{D}}$ is modeled by using Eq. 2 as follows.

$$\boldsymbol{\mu}_{f,\tau|\boldsymbol{\theta},\boldsymbol{D}} = \mathbb{E}[\boldsymbol{x}_{f,\tau}|\boldsymbol{\theta},\boldsymbol{D}] = \sum_{j=1}^{L_d} d_{f,\tau-j+1}\mathbb{E}[\boldsymbol{b}_{j,f,\tau}|\boldsymbol{\theta},\boldsymbol{D}], \tag{9}$$

where $\mathbb{E}[\cdot|\boldsymbol{\theta},\boldsymbol{D}]$ is an operator for conditional expectation with given $\boldsymbol{\theta}$ and $\boldsymbol{D}$, $\mathbb{E}[s_{i,f,\tau}|\boldsymbol{\theta},\boldsymbol{D}]$ is assumed to be 0 under the independence assumption and 0-mean assumption of each source. On contrary to this, $\mathbb{E}[d_{f,\tau}|\boldsymbol{\theta},\boldsymbol{D}] = d_{f,\tau}$. The multichannel covariance matrix of the microphone input signal can also be expanded as follows:

$$\begin{aligned} \boldsymbol{R}_{x,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} &= \mathbb{E}[(\boldsymbol{x}_{f,\tau}-\boldsymbol{\mu}_{f,\tau|\boldsymbol{\theta},\boldsymbol{D}})(\boldsymbol{x}_{f,\tau}-\boldsymbol{\mu}_{f,\tau|\boldsymbol{\theta},\boldsymbol{D}})^H|\boldsymbol{\theta},\boldsymbol{D}] \\ &= \sum_{i=1}^{N_s} v_{i,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}\boldsymbol{R}_{a_i,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} \\ &+ \sum_{j=1}^{L_d}|d_{f,\tau-j+1}|^2\boldsymbol{R}_{b_j,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}, \end{aligned} \tag{10}$$

where

$$\begin{aligned} v_{i,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} &= \mathbb{E}[|s_{i,f,\tau}|^2|\boldsymbol{\theta},\boldsymbol{D}] \\ \boldsymbol{R}_{a_i,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} &= \mathbb{E}[\boldsymbol{a}_{i,f,\tau}\boldsymbol{a}^H_{i,f,\tau}|\boldsymbol{\theta},\boldsymbol{D}], \\ \boldsymbol{R}_{b_j,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} &= \mathbb{E}[(\boldsymbol{b}_{j,f,\tau}-\hat{\boldsymbol{b}}_{j,f,\tau|\boldsymbol{\theta},\boldsymbol{D}})(\boldsymbol{b}_{j,f,\tau}-\hat{\boldsymbol{b}}_{j,f,\tau|\boldsymbol{\theta},\boldsymbol{D}})^H|\boldsymbol{\theta},\boldsymbol{D}]. \end{aligned}$$

The average vector of echo transfer function $\hat{\boldsymbol{b}}_{j,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}$ is defined as $\mathbb{E}[\boldsymbol{b}_{j,f,\tau}|\boldsymbol{\theta},\boldsymbol{D}]$, which is corresponding to the linear part. Therefore, $\hat{\boldsymbol{b}}_{j,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}$ can be approximated as time-invariant value $\hat{\boldsymbol{b}}_{j,f|\boldsymbol{\theta},\boldsymbol{D}}$.

According to time-invariant approximation of the multichannel covariance matrix by Duong et al., [7] the multichannel covariance matrices are approximated as follows:

$$\begin{aligned} \boldsymbol{R}_{a_i,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} &= \boldsymbol{R}_{a_i,f|\boldsymbol{\theta},\boldsymbol{D}}, \\ \boldsymbol{R}_{b_j,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} &= \boldsymbol{R}_{b_j,f|\boldsymbol{\theta},\boldsymbol{D}},. \end{aligned}$$

3.3 EM algorithm for estimating model parameter

Similarly to the original Duong method, the proposed method uses EM algorithm [11] by estimating the covariance matrix and the average value of the acoustic transfer function. The model parameter

$\boldsymbol{\theta}$ is defined as $\{\{\boldsymbol{R}_{a_i,f|\boldsymbol{\theta},\boldsymbol{D}}\},\{\boldsymbol{R}_{b_j,f|\boldsymbol{\theta},\boldsymbol{D}}\},\{v_{i,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}\},\{\hat{\boldsymbol{b}}_{j,f|\boldsymbol{\theta},\boldsymbol{D}}\}\}$. In E step, sufficient statistics of $\boldsymbol{c}_{i,f,\tau}$, $\boldsymbol{e}_{j,f,\tau}$ under the given parameter $\boldsymbol{\theta}_t$ are obtained. In M step, new parameter vector $\boldsymbol{\theta}_{t+1}$ is obtained by maximizing the likelihood function by using the sufficient statistics that are obtained in the E step.

The E step and the M step can be described as follows:

E step:

$$\boldsymbol{W}_{c,i} = v_{i,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}}\boldsymbol{R}_{a_i,f|\boldsymbol{\theta}_t,\boldsymbol{D}}\boldsymbol{R}^{-1}_{x,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}} \tag{11}$$

$$\boldsymbol{W}_{e,j} = |d_{f,\tau-j+1}|^2\boldsymbol{R}_{b_j,f|\boldsymbol{\theta}_t,\boldsymbol{D}}\boldsymbol{R}^{-1}_{x,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}} \tag{12}$$

$$\hat{\boldsymbol{c}}_{i,f,\tau} = \boldsymbol{W}_{c,i}(\boldsymbol{x}_{f,\tau} - \sum_{j_2} d_{f,\tau-j_2+1}\hat{\boldsymbol{b}}_{j_2,f|\boldsymbol{\theta}_t,\boldsymbol{D}}) \tag{13}$$

$$\hat{\boldsymbol{e}}_{j,f,\tau} = \boldsymbol{W}_{e,j}(\boldsymbol{x}_{f,\tau} - \sum_{j_2} d_{f,\tau-j_2+1}\hat{\boldsymbol{b}}_{j_2,f|\boldsymbol{\theta}_t,\boldsymbol{D}}) \tag{14}$$

$$\hat{\boldsymbol{R}}_{c_i,f|\boldsymbol{\theta}_t,\boldsymbol{D}} = \hat{\boldsymbol{c}}_{i,f,\tau}\hat{\boldsymbol{c}}^H_{i,f,\tau} + v_{i,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}}(\boldsymbol{I}-\boldsymbol{W}_{c,i})\boldsymbol{R}_{a_i,f|\boldsymbol{\theta}_t,\boldsymbol{D}}, \tag{15}$$

$$\hat{\boldsymbol{R}}_{e_j,f|\boldsymbol{\theta}_t,\boldsymbol{D}} = \hat{\boldsymbol{e}}_{j,f,\tau}\hat{\boldsymbol{e}}^H_{j,f,\tau} + |d_{f,\tau-j+1}|^2(\boldsymbol{I}-\boldsymbol{W}_{e,j})\boldsymbol{R}_{b_j,f|\boldsymbol{\theta}_t,\boldsymbol{D}}, \tag{16}$$

where $\boldsymbol{I}$ is an identity matrix.

M step:

The model parameter vector $\boldsymbol{\theta}_{t+1}$ is obtained in the M step as follows:

$$v_{i,f,\tau|\boldsymbol{\theta}_{t+1},\boldsymbol{D}} = \frac{1}{M}\mathrm{tr}\Big(\boldsymbol{R}^{-1}_{a_i,f|\boldsymbol{\theta}_t,\boldsymbol{D}}\hat{\boldsymbol{R}}_{c_i,f|\boldsymbol{\theta}_t,\boldsymbol{D}}\Big), \tag{17}$$

$$\boldsymbol{R}_{a_i,f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}} = \frac{1}{L}\sum_{\tau}\frac{1}{v_{i,f,\tau|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}}\hat{\boldsymbol{R}}_{c_i,f|\boldsymbol{\theta}_t,\boldsymbol{D}}, \tag{18}$$

$$\boldsymbol{R}_{b_j,f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}} = \frac{1}{L}\sum_{\tau}\frac{1}{|d_{f,\tau-j+1}|^2}\hat{\boldsymbol{R}}_{b_j,f|\boldsymbol{\theta}_t,\boldsymbol{D}}, \tag{19}$$

$\hat{\boldsymbol{b}}_{j,f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}$ is obtained so as to maximize the original likelihood function $L(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{D})$ under the condition that $\boldsymbol{R}_{x,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} = \boldsymbol{R}_{x,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}}$ as follows:

$$\begin{aligned}
&\frac{\partial L(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{D},\boldsymbol{R}_{x,f,\tau|\boldsymbol{\theta},\boldsymbol{D}} = \boldsymbol{R}_{x,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}})}{\partial \boldsymbol{b}_{j,f|\boldsymbol{\theta},\boldsymbol{D}}}\Big|_{\hat{\boldsymbol{b}}_{j,f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}} = 0 \\
&\leftrightarrow \sum_{\tau}\boldsymbol{R}^{-1}_{x,f,\tau|\theta_t,\boldsymbol{D}}\boldsymbol{B}_{f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}\boldsymbol{D}_{f,\tau}\boldsymbol{D}^H_{f,\tau} = \sum_{\tau}\boldsymbol{R}^{-1}_{x,f,\tau|\theta_t,\boldsymbol{D}}\boldsymbol{x}_{f,\tau}\boldsymbol{D}^H_{f,\tau} \\
&\leftrightarrow \mathrm{vec}\Big(\boldsymbol{B}_{f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}\Big) = \Big(\sum_{\tau}(\boldsymbol{D}_{f,\tau}\boldsymbol{D}^H_{f,\tau})^T \otimes \boldsymbol{R}^{-1}_{x,f,\tau|\theta_t,\boldsymbol{D}}\Big)^{-1}\cdot \\
&\mathrm{vec}\Big(\sum_{\tau}\boldsymbol{R}^{-1}_{x,f,\tau|\theta_t,\boldsymbol{D}}\boldsymbol{x}_{f,\tau}\boldsymbol{D}^H_{f,\tau}\Big),
\end{aligned} \tag{20}$$

where T is the transpose operator of a matrix/vector, vec is a vec operator [12] which expands a matrix into a vector, $\otimes$ is kronecker product of two matrices, $\boldsymbol{B}_{f|\boldsymbol{\theta}_{t+1}}$ is $[\ \hat{\boldsymbol{b}}_{1,f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}\ \cdots\ \hat{\boldsymbol{b}}_{L_d,f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}\ ]$, and $\boldsymbol{D}_{f,\tau}$ is $[\ d_{f,\tau}\ \ \cdots\ \ d_{f,\tau-L_d+1}\ ]^T$.

The proposed method performs the E step and the M step iteratively. The final output signal is MMSE solution $\hat{\boldsymbol{c}}_{i,f,\tau}$ which is defined in Eq. 13. Inter-frequency permutation problem is solved by using power spectral correlation method [13].

3.4 Interpretation of MMSE of near-end speech $\hat{\boldsymbol{c}}_{i,f,\tau}$

Eq. 13 can be divided into the linear echo canceller part, the spatial beamforming part, and the nonlinear echo suppressor part as follows:

$$\begin{aligned}
\hat{\boldsymbol{c}}_{i,f,\tau} \quad = \quad & \underbrace{\frac{v_{i,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}}\sigma_{a,f}}{P_{x,f,\tau}}}_{\text{Non-Linear echo suppressor part}} \\
& \underbrace{\frac{\boldsymbol{R}_{a_i,f|\boldsymbol{\theta}_t,\boldsymbol{D}}}{\sigma_{a,f}}\Big(\frac{\boldsymbol{R}_{x,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}}}{P_{x,f,\tau}}\Big)^{-1}}_{\text{Spatial beamformer}} \\
& \underbrace{(\boldsymbol{x}_{f,\tau} - \sum_{j_2} d_{f,\tau-j_2+1}\hat{\boldsymbol{b}}_{j_2,f|\boldsymbol{\theta}_t,\boldsymbol{D}})}_{\text{Linear echo canceller part}},
\end{aligned} \tag{21}$$

where $P_{x,f,\tau} = \sum_i v_{i,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}}\sigma_{a,f} + \sum_j |d_{f,\tau-j+1}|^2\sigma_{e,f}$, $\sigma_{a,f} = \mathrm{tr}(\boldsymbol{R}_{a_i,f|\boldsymbol{\theta}_t,\boldsymbol{D}})$, and $\sigma_{e,f} = \mathrm{tr}(\boldsymbol{R}_{b_j,f|\boldsymbol{\theta}_t,\boldsymbol{D}})$. $P_{x,f,\tau}$ is the estimated momentary power of the microphone input signal at the frequency f and the frame τ. In a similar way, $v_{i,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}}\sigma_{a,f}$ is the estimated momentary power of the i-th near-end speech source, and $|d_{f,\tau-j+1}|^2\sigma_{e,f}$ is the estimated momentary power of the j-th tap of the residual echo component. The first part can be regarded as a single channel noise suppressor based on an estimation of signal-to-input ratio at each time-frequency point. The second part is a spatial beamforming from the multichannel signal. The third part is a linear echo canceller in which echo component is suppressed independently at each microphone. These three parts are simultaneously optimized by maximizing the likelihood function. Therefore, the proposed method can be regarded as an optimum coupling of linear echo canceller, multichannel spatial beamformer, and nonlinear echo suppressor.

4. EVALUATION

The proposed method is evaluated from two points of view. The first evaluation is robustness evaluation against phase fluctuation of loudspeaker acoustic transfer function. The second evaluation is echo reduction performance evaluation at a real meeting room. The common parameters are listed in Table 1.

Table 1: Experimental common parameters

Sampling rate [Hz]	8000
Frame size [pt]	512
Frame shift [pt]	128

The proposed method is compared with three approaches. The first approach is linear echo canceller. The second approach is semi-blind LGM with 0-mean and full-rank covariance matrix. The third approach is cascade method of linear echo canceller and full-rank method which is defined in the following subsection.

4.1 An alternative way by cascade method of linear echo canceller and full-rank method

An alternative way to the proposed method is a cascade method of linear echo canceller and 0-mean and full-rank covariance semi-blind LGM. In this approach, estimation of linear part of acoustic echo path $\hat{\boldsymbol{b}}_{j,f|\boldsymbol{\theta}_t,\boldsymbol{D}}$ by Eq. 20 is replaced by the simple least-squares solution as follows:

$$\begin{aligned}
&\mathrm{vec}\Big(\boldsymbol{B}_{f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}\Big) \\
= \quad & \Big(\sum_{\tau}(\boldsymbol{D}_{f,\tau}\boldsymbol{D}^H_{f,\tau})^T \otimes \boldsymbol{I}\Big)^{-1}\mathrm{vec}\Big(\sum_{\tau}\boldsymbol{x}_{f,\tau}\boldsymbol{D}^H_{f,\tau}\Big).
\end{aligned}$$

$\boldsymbol{B}_{f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}$ does not depend on the parameter vector $\boldsymbol{\theta}$. Therefore, $\boldsymbol{B}_{f|\boldsymbol{\theta}_{t+1},\boldsymbol{D}}$ can be estimated before performing the EM algorithm.

4.2 Robustness evaluation against phase fluctuation of loudspeaker acoustic transfer function

Analysis is performed by using the virtual impulse response with impulse response generator [14]. The number of microphones was set to be 3. A equilateral triangle array whose one-side is 4 cm was used. Reverberation time was set to be 0.2 [sec]. The number of near-end sources is set to be 3. In this evaluation, the phase component of the loudspeaker signal is modified artificially as follows:

$$d_{f,\tau} = |d_{f,\tau}| \exp(\phi_{d_{f,\tau}} + \pi p_{f,\tau}), \tag{22}$$

where $\phi_{d_{f,\tau}}$ is the phase component of $d_{f,\tau}$, $p_{f,\tau}$ is a random variable whose distribution is 0-mean and ε^2-variance Gaussian distribution.

The number of EM iterations is set to be 100 in this experiment. The evaluation measure is NRR (Noise Reduction Ratio) [dB] which is defined as follows:

$$\mathrm{NRR} = -10\log_{10}\frac{\sum_t(\sum_i \hat{s}_i(t) - \sum_i s_i(t))^2}{\sum_t(x(t) - \sum_i s_i(t))^2}, \tag{23}$$

where $\hat{s}_i(t)$ is an estimation for $s_i(t)$ (the i-th near-end source signal), $x(t)$ is microphone input signal. The experimental result with variable ε is shown in Fig. 1. It is shown that the proposed method

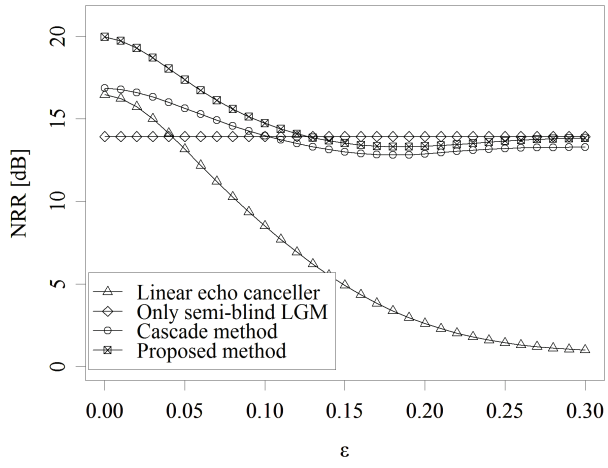


Figure 1: Evaluation result of robustness against phase fluctuation of loudspeaker transfer function

is superior to other methods at wide range of ε. Especially, the proposed method is superior to the cascade method. The linear part of the cascade method is independently optimized from the latter semi-blind LGM part. On contrary to the cascade method, the linear part of the propose method is affected by the multichannel beamforming part and the nonlinear echo suppressor part through the estimated covariance matrix of the microphone input signal $\boldsymbol{R}_{x,f,\tau|\boldsymbol{\theta},\boldsymbol{D}}$. Intuitively, when the near-end speech signals and the non-linear residual echo is dominant, the weight of the corresponding time-frame is low for estimation of the linear part of the proposed method. The linear part of the proposed method is estimated by using only linear-echo dominant frames.

When ε is small-valued, the linear part of the loudspeaker transfer function is dominant, and the linear echo canceller is superior to the semi-blind LGM with 0-mean and full-rank covariance. In the proposed method, the linear echo canceller part is automatically used on a priority basis in this case. When the ε is large-valued, the nonlinear part of the loudspeaker transfer function is dominant, and echo reduction performance of the linear echo canceller is dramatically decreased. In this case, the semi-blind LGM with 0-mean and full-rank covariance is shown to be effective, and the proposed method automatically selects the multichannel beamforming part and the nonlinear echo suppressor part. It is clear by analyzing spatial beam-pattern of the proposed method. Spatial beampattern of the proposed method with various ε is shown in Fig. 2. It is said that the spatial null is steeper with increase of the effect of the fluctuation and decrease of echo reduction performance of the linear echo canceller. In the proposed method, when the phase fluctuation

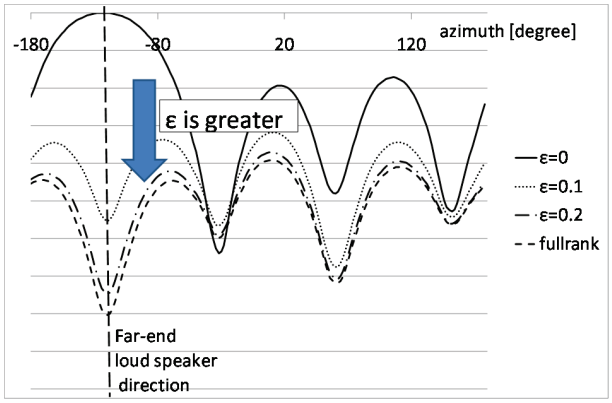


Figure 2: Beampattern of the proposed method

of loudspeaker acoustic transfer function is large, the spatial beamforming part reduces the acoustic echo component instead of the linear echo canceller part.

Next, the proposed method under the condition that the number of near-end speech sources is set to be 3 is compared with the proposed method under the condition that the number of near-end speech sources is set to be 1 (single source model). The evaluation result is shown in Fig. 3. The proposed method under the condition

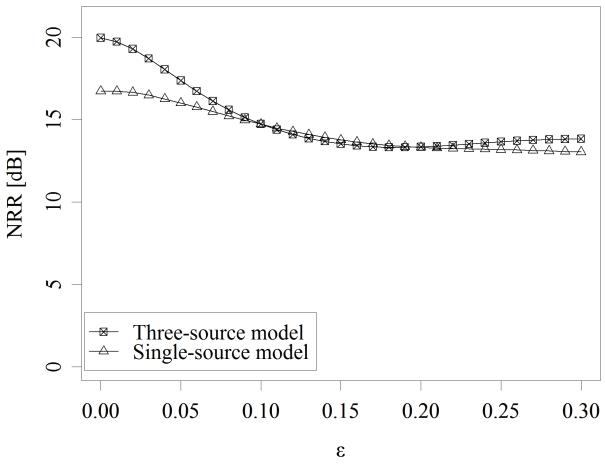


Figure 3: Comparison of proposed method with single source model

that the number of near-end speech sources is 3 is better than the single source model. The time-variant coefficient, $v_{i,f,\tau|\boldsymbol{\theta}_t,\boldsymbol{D}}\sigma_{a,f}$, is not correctly estimated in the single source model. The estimation error of the time-variant coefficient affects the estimation error of $\boldsymbol{R}_x$. Therefore, the linear echo canceller part is correctly estimated in the single source model.

4.3 Echo reduction performance evaluation at a real meeting room

The proposed method is evaluated by using impulse responses which are measured in the real meeting room. The room is shown in Fig. 4. The reverberation time is about 500 ms. The recording sampling rate was 32000 Hz. Impulse responses were downsampled to 8000 Hz. The number of near-end sources is set to be 2. The distance between a microphone array and sources is 1 m. The

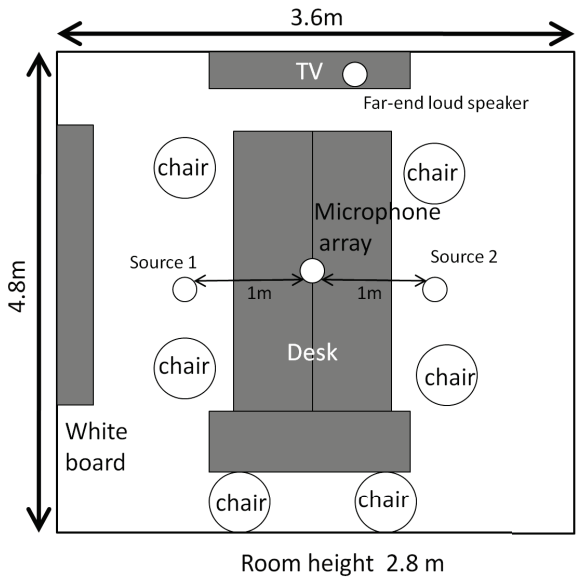
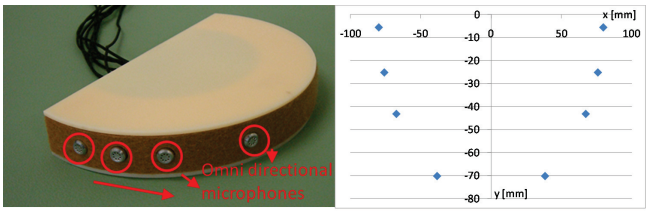


Figure 4: Experimental environment

number of the microphones is set to be 8. The microphone array is shown in Fig. 5. The microphone elements are omni-directional. The number of EM iterations is set to be 100 in this experiment.



(a) Appearance of microphone array (b) Position of microphones

Figure 5: Microphone array used for evaluation

NRR of each source is evaluated in this experiment. Experimental results are shown in Table. 2.

Table 2: Experimental result in a real meeting room: evaluation measure is NRR [dB]

	Semi-blind LGM with 0-mean and full-rank covariance matrix	Cascade method	Proposed method
Average	9.96	10.50	**11.00**
Source 1	10.20	10.80	**11.09**
Source 2	9.72	10.19	**10.92**

The proposed method is superior to "Semi-blind LGM with 0-mean and full-rank covariance matrix" and "Cascade method". "Semi-blind LGM with 0-mean and full-rank covariance matrix" is inferior to the proposed method, because cross spectral term of the far-end loudspeaker signals is neglected in this method.

5. CONCLUSION AND FUTURE WORKS

In this paper, a novel echo reduction technique was proposed. The proposed method optimizes linear echo canceller part, multichannel spatial beamforming part, and nonlinear echo suppressor part. These three parts are optimized by using the same likelihood function based on semi-blind local Gaussian modeling (semi-blind LGM). All three parts are affected by each other. This is the major difference between the proposed method and a cascade method of the linear echo canceller and semi-blind local Gaussian model with 0-mean and full-rank covariance model. The experimental result show that the proposed method is superior to the linear echo canceller, semi-blind LGM with 0-mean and full-rank covariance model, and the cascade model.

The proposed method is an *offline* method. In the real TV conferencing system, the optimization scheme which can be used for sample-by-sample execution. The proposed method can be easily extended to an *online* method by using an online LGM scheme proposed by authors [8] in the near future.

REFERENCES

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

[2] O. Hoshuyama and A. Sugiyama, "An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo," *Proc. ICASSP2006*, vol. V, pp. 269–272, 2006.

[3] P.C. Loizou, "Speech Enahcement: Theory And Practice," Signal Processing and Communications, Crc Pr Llc, 2007/6.

[4] G. Reuven, S.Gannot, I. Cohen, "Multichannel acoustic echo cancellation and noise reduction in reverberant environments using the transfer-function GSC," *Proc. IEEE ICASSP2007*, 2007.

[5] Q. Lin, N Xu, and H. Liang, "A semi-blind EM algorithm for overcomplete ICA, " *IEEE ICASSP2009*, 2009.

[6] E. Vincent et al., "Underdetermined instantaneous audio source separation via local Gaussian modeling," *Proc. ICA '09*, pp. 775–782, 2009.

[7] N.Q.k. Duong et al., "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010/9.

[8] M. Togami, "Online speech source separation based on maximum likelihood of Local Gaussian modeling," *Proc. ICASSP2011*, pp. 213–216, 2011/5.

[9] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation," In *Proc. ICASSP2009*, pp. 3137–3140, 2009.

[10] M. Togami, Y. Kawaguchi, H. Kokubo, and Y. Obuchi, "Acoustic echo suppressor with multichannel semi-blind nonnegative matrix factorization, " *Proc. APSIPA 2010*, 2010/12.

[11] 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.

[12] D.A. Harville, *Matrix Algebra from a Statistician's Perspective.* New York: Springer-Verlag, 1997.

[13] S. Ikeda and N. Murata, "An approach to blind source separation of speech signals," In *Proc. ICANN '98,* pp. 761–766, 1998.

[14] Room Impulse Response Generator for MATLAB, http://home.tiscali.nl/enabets/rir_generator.html.