Over-Determined Semi-Blind Speech Source Separation

Masahito Togami* and Robin Scheibler*
* LINE Corporation, Tokyo, Japan
E-mail: masahito.togami@linecorp.com

***Abstract*—We propose a semi-blind speech source separation that jointly optimizes several acoustic functions, i.e., speech source separation (SS), dereverberation (DR), acoustic echo reduction (AE), and background noise reduction (BG). Instead of cascade connection of SS, AE, and DR, the proposed method performs DR, SS, and AE by a unified time-invariant filter. We assume the over-determined condition that the number of the microphones N_m is larger than the number of near-end speech sources N_s. Joint optimization of DR, SS, AE, and BG can be performed by using the $N_m - N_s$ dimensional subspace of the time-invariant filter for BG. Furthermore, we reveal that this subspace can be also utilized for residual acoustic echo reduction (RR) in which residual acoustic echo signal is reduced by spatial filtering. Two types of joint parameter optimization techniques for DR, SS, AE, BG, and RR are proposed based on vectorwise coordinate descent and fast multichannel nonnegative matrix factorization. Experimental results show that the proposed methods perform DR, SS, AE, and BG better than a cascade method. When the acoustic echo path between a loudspeaker and a microphone is time-varying, performance improvement of the proposed methods with RR is larger than the proposed method without RR.**

Index Terms: acoustic echo reduction, speech source separation, dereverberation, noise reduction

I. Introduction

Speech source separation is an important technique for human-listening systems and automatic speech recognition (ASR) systems. Speech enhancement and blind speech source separation techniques have been actively studied [1]–[4]. Dereverberation techniques [5], [6] are also important techniques for speech applications that are utilized under reverberant environments. An acoustic echo signal which is an output signal of a loudspeaker is also an unwanted signal in teleconferencing systems and AI Assistants. Thus, speech source separation (SS), dereverberation (DR), background noise reduction (BG), and acoustic echo reduction (AE) have been actively studied for a long time. Cascade connection of several acoustic functions is one of the solutions. However, the output signal is not optimized. For example, when AE is performed prior to SS and/or BG [7], a time-invariant filter of AE is poorly optimized due to the existence of the other speech sources. Thus, joint optimization of several acoustic functions has been strongly required.

As joint optimization of several acoustic functions, multichannel local Gaussian modeling (LGM) [8] based approaches have been considered [9]–[11]. Joint optimization of SS, DR, BG, and AE outperformed cascade-connection based methods [11]. In this method, the residual acoustic echo signal after AE is also reduced by spatial filtering. It is highly effective to prevent from howling of teleconferencing systems and to remove an extremely large acoustic echo signal when a loudspeaker is attached closely to a microphone. However, the computational cost is too high due to the calculation of an inverse matrix whose dimension is proportional to the square of the number of the microphones and the tap-length of a time-invariant filter. Because this matrix is an output matrix of Kronecker product, we call this problem the Kronecker product problem.

Recently, determined speech source separation techniques have been actively studied [12]–[17]. Determined speech source separation assumes that the number of microphones N_m is equal to the number of speech sources N_s. Determined speech source separation such as independent low-rank matrix analysis (ILRMA) [13] is known to be more stable than the LGM based approaches. Joint optimization of SS and DR has been proposed [14], [16]. Kagami et al. [14] optimize a dereverberation filter similarly to the LGM based method [10], and the Kronecker product problem is also problematic in this method. On the other hand, the recently proposed ILRMA-T [16] does not require Kronecker product, and computational cost is lower than [14]. As an extension of ILRMA-T, joint optimization of SS, DR, and BG has been also proposed [18], [19] (OverILRMA-T). In the OverILRMA-T, N_m is assumed to be larger than N_s. The OverILRMA-T is based on a determined speech source separation with a time-invariant filter. Thus, the N_s dimensional subspace of the time-invariant filter is sufficient for SS. The remained $N_m - N_s$ dimensional subspace is utilized for BG [20]–[22].

In this paper, we propose a joint optimization of SS, DR, BG, and AE. At first, it is shown that a time-invariant filter for AE can be naturally integrated with a time-invariant filter for SS and DR. A unified time-invariant filter is optimized similarly to the ILRMA-T framework. Thus, the Kronecker product problem is not problematic in this framework. Similarly to the OverILRMA-T, the proposed method assumes that the number of microphones N_m is larger than the number of near-end speech sources N_s. BG is performed by using the $N_m - N_s$ dimensional subspace. We call this framework OverILRMA-T-AE. We further propose the utilization of the $N_m - N_s$ dimensional subspace for not only BG but also residual acoustic echo reduction (RR), similarly to the LGM based approaches [9], [11]. When an acoustic impulse response

between a loudspeaker and a microphone is time-varying, the acoustic echo signal is not sufficiently removed by only the time-invariant filter. The proposed method reduces the acoustic signal in two ways, i.e., a time-invariant filter and a time-varying multi-channel spatial filter. We propose two parameter-optimization techniques for joint optimization. The first one is an extension of vectorwise coordinate descent (VCD) [23] based approach. The second one is an extension of fast multi-channel nonnegative matrix factorization (FMNMF) approach with joint diagonalization [24]. Experimental results show that the proposed OverILRMA-T-AE framework outperformed the cascade connection of AE and OverILRMA-T. It is also shown that the proposed joint optimization of DR, SS, AE, BG, and RR with VCD and FMNMF outperformed OverILRMA-T-AE especially when the acoustic echo path is time-varying. From the computational cost perspective, the proposed method with FMNMF is shown to be more efficient than the proposed method with VCD.

II. Modeling

A. Microphone input signal

Speech source separation is performed in a time-frequency domain. Multi-channel microphone input signal $\boldsymbol{x}_{lk}$ (l is the frame index, k is the frequency index) is modeled in the time-frequency domain as follows:

$$\boldsymbol{x}_{lk} = \boldsymbol{A}_k \boldsymbol{s}_{lk} + \boldsymbol{e}_{lk} + \boldsymbol{r}_{lk} + \boldsymbol{n}_{lk}, \tag{1}$$

where $\boldsymbol{x}_{lk} \in \mathbb{C}^{N_m}$, N_m is the number of the microphones, $\boldsymbol{s}_{lk} \in \mathbb{C}^{N_s}$ is a vector which contains near-end speech source signals, N_s is the number of the near-end speech sources, $\boldsymbol{A}_k$ is a matrix which contains the steering vector of each speech source, $\boldsymbol{e}_{lk}$ is the spatial image of the acoustic echo signal, $\boldsymbol{r}_{lk}$ is the late reverberation term, and $\boldsymbol{n}_{lk}$ is the background noise term. N_m is assumed to be larger than N_s. $\boldsymbol{e}_{lk}$ is modeled as follows:

$$\boldsymbol{e}_{lk} = \sum_{\tau=0}^{L_e-1} \boldsymbol{g}_{\tau k} d_{l-\tau,k}, \tag{2}$$

where L_e is the tap-length of the acoustic echo path, $\boldsymbol{g}$ is the impulse response of the acoustic echo path, and d is the pre-given original signal. The reverberation term $\boldsymbol{r}_{lk}$ is modeled with an autoregressive model [6], [25] as follows:

$$\boldsymbol{r}_{lk} = \sum_{\tau=0}^{L_r-1} \boldsymbol{F}_{\tau k} \boldsymbol{x}_{l-\tau,k}, \tag{3}$$

where L_r is the tap-length of the autoregressive coefficient and $\boldsymbol{F}$ is the autoregressive coefficient that estimates late reverberation from the past microphone input signal. $\boldsymbol{e}_{lk}$ and $\boldsymbol{r}_{lk}$ can be combined into one term, and the microphone input signal $\boldsymbol{x}_{lk}$ is re-modeled as follows:

$$\boldsymbol{x}_{lk} = \boldsymbol{A}_k \boldsymbol{s}_{lk} + \boldsymbol{G}_k \tilde{\boldsymbol{x}}_{lk} + \boldsymbol{n}_{lk}, \tag{4}$$

where

$$\tilde{\boldsymbol{x}}_{lk} = \begin{bmatrix} \boldsymbol{d}_{lk}^T & \bar{\boldsymbol{x}}_{lk}^T \end{bmatrix}^T, \tag{5}$$

$$\boldsymbol{d}_{lk} = \begin{bmatrix} d_{l,k} & \cdots & d_{l-L_e+1,k} \end{bmatrix}^T, \tag{6}$$

$$\bar{\boldsymbol{x}}_{lk} = \begin{bmatrix} \boldsymbol{x}_{l-D,k}^T & \cdots & \boldsymbol{x}_{l-D-L_r+1,k}^T \end{bmatrix}^T, \tag{7}$$

and T is the transpose operator of a matrix/vector. The objective of speech source separation is defined as the extraction of the spatial image of each speech source $s_{ilk}\boldsymbol{a}_{ik}$ ($\boldsymbol{a}_{ik}$ is the i-th column vector of $\boldsymbol{A}_k$) from $\boldsymbol{x}_{lk}$ defined in (4). Because d is known in advance, this speech source separation problem can be interpreted as a semi-blind speech source separation problem.

B. Probabilistic modeling

1) Overview: Speech source separation is performed in a probabilistic way. In this paper, an over-determined model [20]–[22] is introduced. In the over-determined model, the spatial covariance matrix (SCM) of each source is modeled as a rank-one matrix. The SCM of the residual signal is modeled as a $N_m - N_s$ dimensional matrix. Summation of the dimensions of the speech SCMs and the dimension of the residual SCM are N_m. Thus, speech source separation is performed with time-invariant multi-channel filtering.

2) Speech source model and residual signal model: Each speech source s_{ilk} is modeled as the following time-varying Gaussian distribution [13]–[17]:

$$p\left(s_{ilk}\right) = \mathcal{N}\left(0, v_{ilk}\right), \tag{8}$$

where v_{ilk} is the time-varying variance of the i-th speech source, which is modeled based on the non-negative matrix factorization (NMF) as follows:

$$v_{ilk} = \sum_{n=1}^{N_n} c_{iln} b_{ink}, \tag{9}$$

N_n is the number of basis vectors, $b_{ink} \geq 0$ is the basis coefficient of the n-th component, and $c_{iln} \geq 0$ is the time-varying activity of the n-th component.

The residual signal $\boldsymbol{r}_{lk} = \boldsymbol{G}_k \tilde{\boldsymbol{x}}_{lk} + \boldsymbol{n}_{lk}$ is modeled as the following $N_m - N_s$ dimensional time-varying Gaussian distribution:

$$p\left(\boldsymbol{r}_{lk}\right) = \mathcal{N}\left(\boldsymbol{G}_k \tilde{\boldsymbol{x}}_{lk}, \boldsymbol{V}_{lk}\right), \tag{10}$$

where $\boldsymbol{V}_{lk}$ is a $N_m - N_s$ dimensional covariance matrix of the residual signal.

3) Probabilistic model of microphone input signal: The microphone input signal is modeled as a time-varying Gaussian distribution, because all components, i.e., speech sources and residual signal, are modeled as time-varying Gaussian distributions. The time-varying Gaussian distribution of the microphone input signal is modeled as follows:

$$p\left(\boldsymbol{x}_{lk}\right) = \mathcal{N}\left(\boldsymbol{G}_k \tilde{\boldsymbol{x}}_{lk}, \boldsymbol{R}_{xlk}\right), \tag{11}$$

where $\boldsymbol{R}_{xlk}$ is the following time-varying covariance matrix of the microphone input signal:

$$\boldsymbol{R}_{xlk} = \overline{\boldsymbol{A}}_k \mathrm{diag}\begin{pmatrix} v_{1lk} & \cdots & v_{N_s lk} & \boldsymbol{V}_{lk} \end{pmatrix} \overline{\boldsymbol{A}}_k^H, \tag{12}$$

where H is the Hermitian transpose of a matrix/vector, and $\overline{\boldsymbol{A}}_k = \begin{pmatrix} \boldsymbol{A}_k & \boldsymbol{A}_{nk} \end{pmatrix}$. Finally, the negative log-likelihood function of the microphone input signal $\mathcal{L}_{lk} = -\log p\left(\boldsymbol{x}_{lk}\right)$ can be modeled as follows:

$$\begin{aligned} \mathcal{L}_{lk} = \sum_{i=1}^{N_s} \frac{\left|\boldsymbol{p}_{ik}^H \tilde{\boldsymbol{x}}_{lk}\right|^2}{\sum_{n=1}^{N_n} c_{iln} b_{ink}} + \left(\boldsymbol{P}_{nk} \tilde{\boldsymbol{x}}_{lk}\right)^H \boldsymbol{V}_{lk}^{-1} \boldsymbol{P}_{nk} \tilde{\boldsymbol{x}}_{lk} \\ + \log \left|\det \boldsymbol{V}_{lk}\right| - 2 \log \left|\det \boldsymbol{W}_k\right|, \end{aligned} \tag{13}$$

where $\boldsymbol{W}_k = \overline{\boldsymbol{A}}_k^{-1}$ and

$$\begin{aligned} \boldsymbol{P}_k &= \begin{bmatrix} \boldsymbol{W}_k & -\boldsymbol{W}_k \boldsymbol{G}_k \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{p}_{1k} & \cdots & \boldsymbol{p}_{N_s k} & \boldsymbol{P}_{nk}^H \end{bmatrix}^H. \end{aligned} \tag{14}$$

$\boldsymbol{p}_{ik}$ is a unified time-invariant filter which extracts the i-th speech source by performing DR, SS, and AE jointly.

III. Proposed method

All parameters are optimized to minimize $\mathcal{L} = \sum_{lk} \mathcal{L}_{lk}$. However, there is no closed-form solution for parameter optimization. The parameters are updated based on the majorization-minimization (MM) algorithm [26] iteratively. A monotonical decrease of the cost function is assured in each iteration.

A. Parameter optimization of NMF parameters

NMF parameters are updated to minimize the cost function $\mathcal{L}$ in an iterative manner as follows:

$$b_{ink} \leftarrow b_{ink} \sqrt{\frac{\sum_t \left|\hat{s}_{ilk}\right|^2 c_{iln} \left(\sum_{n=1}^{N_n} c_{iln} b_{ink}\right)^{-2}}{\sum_t c_{iln} \left(\sum_{n=1}^{N_n} c_{iln} b_{ink}\right)^{-1}}}, \tag{15}$$

$$c_{iln} \leftarrow c_{iln} \sqrt{\frac{\sum_k \left|\hat{s}_{ilk}\right|^2 b_{ink} \left(\sum_{n=1}^{N_n} c_{iln} b_{ink}\right)^{-2}}{\sum_k b_{ink} \left(\sum_{n=1}^{N_n} c_{iln} b_{ink}\right)^{-1}}}, \tag{16}$$

where $\hat{s}_{ilk}$ is the separated signal defined as $\hat{s}_{ilk} = \boldsymbol{p}_{ik}^H \tilde{\boldsymbol{x}}_{lk}$.

B. Parameter optimization with orthogonal constraint

When the covariance matrix of the residual signal $\boldsymbol{V}_{lk}$ is a time-varying matrix with no constraint, some speech sources in N_s sources are mistakenly assigned in the residual signal term, and these sources are missing in the output signal. In the conventional over-determined model, $\boldsymbol{V}_{lk}$ was set to a time-invariant matrix $\boldsymbol{R}_{nk}$ under the assumption that the residual signal is the stationary background noise. In this case, although $\tilde{\boldsymbol{x}}_{lk}$ contains not only the past microphone input signal but also the acoustic echo signal, $\boldsymbol{P}_k$ can be updated similarly to the conventional over-determined ILRMA-T model (OverILRMA-T) with orthogonal constraint [18], [19]. The first N_s rows of $\boldsymbol{P}_k$ can be updated based on the ILRMA-T based separation filter update [15], [16] as follows:

$$\boldsymbol{p}_{ik} \leftarrow \frac{\boldsymbol{Q}_{ik}^{-1} \boldsymbol{z}_{ik}}{\sqrt{\boldsymbol{z}_{ik}^H \boldsymbol{Q}_{ik}^{-1} \boldsymbol{z}_{ik}}}, \tag{17}$$

where

$$\boldsymbol{Q}_{ik} = \frac{1}{L_T} \sum_{l=1}^{L_T} \frac{\tilde{\boldsymbol{x}}_{lk} \tilde{\boldsymbol{x}}_{lk}^H}{\sum_{n=1}^{N_n} c_{iln} b_{ink}}, \tag{18}$$

$$\boldsymbol{z}_{ik} = \begin{pmatrix} \boldsymbol{W}_k^{-1} \boldsymbol{e}_i \\ \boldsymbol{0} \end{pmatrix}, \tag{19}$$

L_T is the number of time-frames, and $\boldsymbol{e}_i$ takes 1 in only the i-th element and takes 0 in the other elements. The remained term of $\boldsymbol{P}_k$, i.e., $\boldsymbol{P}_{nk}$, is updated with the orthogonal constraint [18], [19] as follows:

$$\boldsymbol{P}_{nk} \leftarrow \begin{pmatrix} \boldsymbol{C}_{nk} \\ -\boldsymbol{I} \\ \boldsymbol{J}_{nk,3} \boldsymbol{J}_{nk,1}^{-1} \boldsymbol{C}_{nk} - \boldsymbol{J}_{nk,3} \boldsymbol{J}_{nk,1}^{-1} \boldsymbol{E}_n \end{pmatrix}, \tag{20}$$

where

$$\begin{pmatrix} \boldsymbol{J}_{nk,1,N_m \times N_m} & \boldsymbol{J}_{nk,2} \\ \boldsymbol{J}_{nk,3} & \boldsymbol{J}_{nk,4} \end{pmatrix} = \boldsymbol{Q}_{nk}^{-1}, \tag{21}$$

$$\boldsymbol{Q}_{nk} = \sum_{l=1}^{L_T} \frac{\tilde{\boldsymbol{x}}_{lk} \tilde{\boldsymbol{x}}_{lk}^H}{L_T}, \tag{22}$$

$$\boldsymbol{C}_{nk} = \left(\boldsymbol{W}_{s,k} \boldsymbol{J}_{nk,1}^{-1} \boldsymbol{E}_s\right)^{-1} \boldsymbol{W}_{s,k} \boldsymbol{J}_{nk,1}^{-1} \boldsymbol{E}_n, \tag{23}$$

and $\boldsymbol{W}_{s,k}$ is a $N_s \times N_m$ dimensional submatrix of $\boldsymbol{W}_k$. The output $\boldsymbol{P}_k$ is optimized to perform DR, SS, AE, and BG jointly. We call this model OverILRMA-T-AE-OC.

C. Parameter optimization with time-varying residual covariance matrix model

In the OverILRMA-T-AE-OC, the acoustic echo signal is reduced by only time-invariant linear filtering. However, when the impulse response between a loudspeaker and a microphone is time-varying, there is a residual acoustic echo signal after the time-invariant linear filtering. We propose a residual acoustic echo reduction (RR) by using multi-channel spatial filtering. We introduce a time-varying covariance matrix which is correlated with the power of the acoustic echo signal so that the residual echo signal is correctly assigned in the residual signal term similarly to LGM based methods [9], [11]. The time-varying covariance matrix is modeled as follows:

$$\boldsymbol{V}_{lk} = \boldsymbol{R}_{nk} + \sum_{\tau=0}^{L_e-1} \left|d_{l-\tau,k}\right|^2 \boldsymbol{U}_{\tau k}, \tag{24}$$

where $\boldsymbol{U}_{\tau k}$ is the multi-channel covariance matrix of the τ-th tap of the residual acoustic echo signal. Even though $\boldsymbol{U}_{\tau k}$ is updated with no constraint, $\left|d_{l-\tau,k}\right|^2 \boldsymbol{U}_{\tau k}$ is not correlated with the power of the N_s speech sources under the assumption that the acoustic echo signal and the N_s speech sources are independent of each other. Thus, it is expected that any speech source is not assigned mistakenly in the residual signal term. For optimizing the proposed time-varying residual covariance matrix model and $\boldsymbol{P}_{nk}$, we propose two optimization algorithms. In each method, each low vector of $\boldsymbol{P}_{nk}$ is updated sequentially.

1) Vectorwise coordinate descent based optimization: We extend the vectorwise coordinate descent (VCD) based optimization algorithm [23] for optimization $\boldsymbol{P}_{nk}$ with the time-varying covariance matrix of the residual signal. When the t-th row vector of $\boldsymbol{P}_{nk}$ is updated, terms which are related to $\boldsymbol{p}_{tk}$ in $\mathcal{L}$ are summarized as follows:

$$\mathcal{F}\left(\boldsymbol{p}_{tk}\right) = \boldsymbol{p}_{tk}^H \boldsymbol{Q}_{ttk} \boldsymbol{p}_{tk} + \boldsymbol{\lambda}_{tk}^H \boldsymbol{p}_{tk} + \boldsymbol{p}_{tk}^H \boldsymbol{\lambda}_{tk} - 2\log\left|\det \boldsymbol{W}_k\right|, \tag{25}$$

where

$$\boldsymbol{\lambda}_{tk} = \frac{\sum_l \sum_{s\neq t} \left(\boldsymbol{V}_{lk}^{-1}\right)_{st} \boldsymbol{x}_{lk} \boldsymbol{x}_{lk}^H \boldsymbol{p}_{sk}}{L_T}, \tag{26}$$

$$\boldsymbol{Q}_{ttk} = \frac{\sum_l \boldsymbol{x}_{lk} \boldsymbol{x}_{lk}^H \left(\boldsymbol{V}_{lk}^{-1}\right)_{tt}}{L_T}, \tag{27}$$

$$\hat{\boldsymbol{a}}_{tk} = \begin{bmatrix} \left(\hat{\boldsymbol{W}}_k^{-1} \boldsymbol{e}_{N_m+t}\right)^T & \boldsymbol{0} \end{bmatrix}^T. \tag{28}$$

$\boldsymbol{p}_{tk}$ which minimizes $\mathcal{F}\left(\boldsymbol{p}_{tk}\right)$ under the condition that the other parameters are fixed is obtained as follows:

$$\boldsymbol{p}_{tk} = \boldsymbol{Q}_{ttk}^{-1} \left(\alpha \hat{\boldsymbol{a}}_{tk} - \boldsymbol{\lambda}_{tk}\right), \tag{29}$$

where α is a variable. α is obtained as follows:

$$\alpha = \begin{cases} \frac{1}{\sqrt{\hat{\boldsymbol{a}}_{tk}^H \boldsymbol{Q}_{ttk}^{-H} \hat{\boldsymbol{a}}_{tk}}} & \text{if } \boldsymbol{\lambda}_{tk}^H \boldsymbol{Q}_{ttk}^{-H} \hat{\boldsymbol{a}}_{tk} = 0 \\ -\beta \left(\boldsymbol{\lambda}_{tk}^H \boldsymbol{Q}_{ttk}^{-H} \hat{\boldsymbol{a}}_{tk}\right)^* & \text{otherwise} \end{cases}. \tag{30}$$

where

$$\beta = \frac{-r + \sqrt{r^2 + 4c}}{2c}, \tag{31}$$

$$r = \left|\boldsymbol{\lambda}_{tk}^H \boldsymbol{Q}_{ttk}^{-H} \hat{\boldsymbol{a}}_{tk}\right|^2, \tag{32}$$

$$c = \left|\boldsymbol{\lambda}_{tk}^H \boldsymbol{Q}_{ttk}^{-H} \hat{\boldsymbol{a}}_{tk}\right|^2 \hat{\boldsymbol{a}}_{tk}^H \boldsymbol{Q}_{ttk}^{-H} \hat{\boldsymbol{a}}_{tk}. \tag{33}$$

Optimization of $\boldsymbol{R}_{nk}$ and $\boldsymbol{U}_{\tau k}$ is done based on multichannel nonnegative matrix factorization (MNMF) [27], [28] under the condition that $\boldsymbol{P}_k$ is fixed and that a $N_m - N_s$ dimensional vector $\boldsymbol{y}_{lk} = \boldsymbol{P}_{nk} \tilde{\boldsymbol{x}}_{lk}$ is regarded as the input signal of the MNMF. We call this algorithm *OverILRMA-T-AE-VCD*.

2) Fast MNMF based optimization: We extend the Fast MNMF based optimization [24] for semi-blind speech source separation. We assume that $\boldsymbol{R}_{nk}$ and $\boldsymbol{U}_{\tau k}$ are jointly diagonalized as follows:

$$\boldsymbol{R}_{nk} = \boldsymbol{D}_k \mathrm{diag}\begin{pmatrix} r_{1k} & \cdots & r_{N_m-N_s k} \end{pmatrix} \boldsymbol{D}_k^H, \tag{34}$$

$$\boldsymbol{U}_{\tau k} = \boldsymbol{D}_k \mathrm{diag}\begin{pmatrix} g_{1\tau k} & \cdots & g_{N_m-N_s \tau k} \end{pmatrix} \boldsymbol{D}_k^H, \tag{35}$$

where

$$\boldsymbol{D}_k \in \mathbb{C}^{N_m - N_s \times N_m - N_s}. \tag{36}$$

The i-th low vector of $\boldsymbol{P}_{nk}$, $\boldsymbol{p}_{N_m+i,k}^H$, is updated as follows:

$$\boldsymbol{p}_{N_m+ik} \leftarrow \frac{\boldsymbol{Q}_{N_m+ik}^{-1} \boldsymbol{z}_{ik}}{\sqrt{\boldsymbol{z}_{ik}^H \boldsymbol{Q}_{N_m+ik}^{-1} \boldsymbol{z}_{ik}}}, \tag{37}$$

TABLE I
AVERAGED TIME [SEC] FOR ONE ITERATION

	$N_m = 3$	4	5
AE+OverILRMA-T	1.74	2.88	4.58
OverILRMA-T-AE-OC	2.60	4.12	5.94
OverILRMA-T-AE-VCD	7.65	13.72	29.53
OverILRMA-T-AE-FMNMF	2.53	5.10	8.14

where

$$\boldsymbol{Q}_{N_m+ik} = \frac{1}{L_T} \sum_{l=1}^{L_T} \frac{\tilde{\boldsymbol{x}}_{lk} \tilde{\boldsymbol{x}}_{lk}^H}{r_{ik} + \sum_\tau \left|d_{l-\tau,k}\right|^2 g_{i\tau k}}, \tag{38}$$

$$\boldsymbol{z}_{ik} = \begin{pmatrix} \boldsymbol{W}_k^{-1} \boldsymbol{e}_{N_m+i} \\ \boldsymbol{0} \end{pmatrix}. \tag{39}$$

r is updated similarly to NMF as follows:

$$r_{ik} \leftarrow r_{ik} \sqrt{\frac{\sum_l \frac{|y_{ilk}|^2}{\left(r_{ik} + \sum_\tau |d_{l-\tau,k}|^2 g_{i\tau k}\right)^2}}{\sum_l \frac{1}{r_{ik} + \sum_\tau |d_{l-\tau,k}|^2 g_{i\tau k}}}}, \tag{40}$$

$$g_{i\tau k} \leftarrow g_{i\tau k} \sqrt{\frac{\sum_l \frac{|d_{l-\tau,k}|^2 |y_{ilk}|^2}{\left(r_{ik} + \sum_\tau |d_{l-\tau,k}|^2 g_{i\tau k}\right)^2}}{\sum_l \frac{|d_{l-\tau,k}|^2}{r_{ik} + \sum_\tau |d_{l-\tau,k}|^2 g_{i\tau k}}}}. \tag{41}$$

We call this algorithm *OverILRMA-T-AE-FMNMF*.

D. Output signal with projection back

The spatial image of the separated signal is estimated with the projection back. The steering matrix $\boldsymbol{A}_k$ is obtained as $\boldsymbol{W}_k^{-1}$. The spatial image of the i-th speech source is estimated as follows:

$$\boldsymbol{c}_{ilk} = s_{ilk} \boldsymbol{a}_{ik}. \tag{42}$$

IV. EXPERIMENT

A. Experimental setup

Performances of DR, SS, AE, BG, and RR were evaluated with simulated data made by `Pyroomacoustics` [29]. The sampling rate was 16000 Hz. The frame size was 1024 pt. The frame shift was 512 pt. L_d was 4. L_e was 4. D was 1. N_n was 2. The number of iterations was 100. N_s were 2. N_m were 3, 4, and 5. The speech sources were extracted from the WSJ1 dataset. The number of the evaluation data was 333. RT_{60} was randomly selected from 0.5 [sec] to 0.8 [sec] for each data. The microphone alignment was also randomly selected. SNR between speech sources was randomly selected from -5 dB to 5 dB. SNR between speech sources and the acoustic echo signal was randomly selected from -10 dB to 0 dB. SNR between speech sources and background noise was randomly selected from 10 dB to 30 dB. The background noise was selected from the CHiME3 dataset [30]. Two acoustic conditions were simulated, i.e., a time-invariant condition and a time-varying condition. In the time-invariant condition, the impulse response of the acoustic echo signal (echo path) was time-invariant. In the time-varying condition, the echo path changed only once during each utterance by changing randomly the location of the loudspeaker up to 0.05 m. The

change-time was randomly selected. Evaluation measures were differences of Cepstrum Distortion (CD), SI-SDR, and SI-SIR between before processing and after processing. The lower value is better in Δ CD. The higher value is better in Δ SI-SDR and Δ SI-SDR.

B. Experimental results

We compared the proposed methods with the cascade connection of AE and OveILRMA-T [18], [19]. In Fig. 1, the experimental results when the impulse responses between a loudspeaker and microphones are time-invariant are shown. In Fig. 2, the experimental results when the echo path is time-varying are shown. The OverILRMA-T-AE-OC outperformed the AE+OverILRMA-T. It can be said that the joint optimization of time-invariant acoustic echo reduction filter in the ILRMA-T framework is effective. The OverILRMA-T-AE-VCD and the OverILRMA-T-AE-FMNMF outperformed the OverILRMA-T-AE-OC. Especially when the echo path is time-varying, the performance improvement of the OverILRMA-T-AE-VCD and the OverILRMA-T-AE-FMNMF is larger. Thus, it can be said that RR in the OverILRMA-T-AE-VCD and the OverILRMA-T-AE-FMNMF is effective. When N_m is larger, the difference between the OverILRMA-T-AE-OC and the OverILRMA-T-AE-VCD or the OverILRMA-T-AE-FMNMF is larger. The OverILRMA-T-AE-VCD and the OverILRMA-T-AE-FMNMF can utilize the excess dimension effectively for RR.

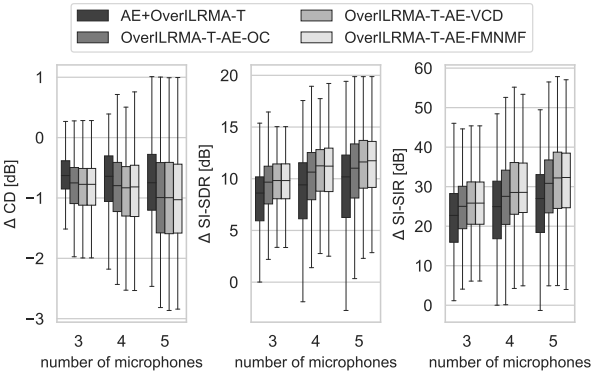

Fig. 1. Box plot when echo path is time-invariant

In Fig. 3 and Fig. 4, convergence speed was evaluated. It is shown that the convergence speed of all methods are approximately equivalent even though maximum performances of the OverILRMA-T-AE-VCD and the OverILRMA-T-AE-FMNMF are higher than those of the AE+OverILRMA-T and the OverILRMA-T-AE-OC. Computation cost was also evaluated in Table I. A server with Intel Xeon Silver 4114 CPU @ 2.20GHz and 128 GB RAM was used. It is shown that OverILRMA-T-AE was slower than AE+OverILRMA-T at the expense of performance improvement. From the comparison between the OverILRMA-T-AE-VCD and the OverILRMA-T-AE-FMNMF, it is shown that by using FMNMF, the computational cost was heavily decreased even though the performance difference between these two methods was not so much.

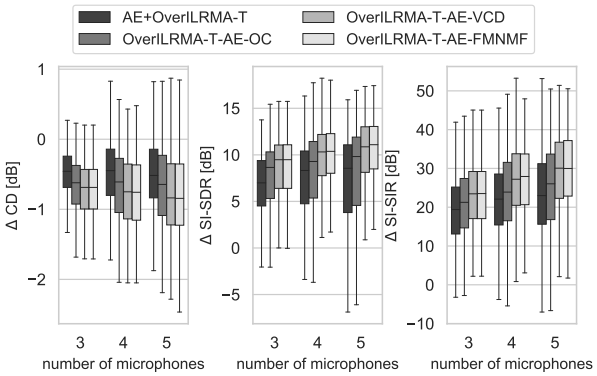

Fig. 2. Box plot when echo path is time-varying

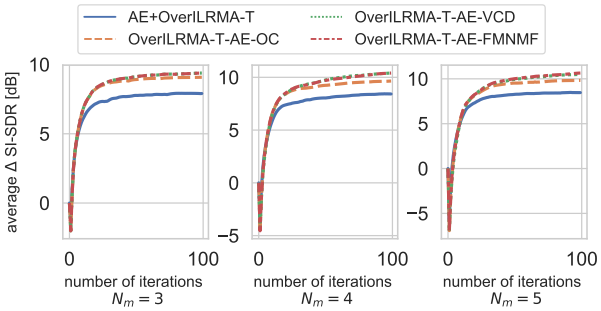

Fig. 3. Convergence speed when echo path is time-invariant

C. Conclusion

We proposed a joint optimization of multi-channel speech source separation, dereverberation, background noise reduction, and acoustic echo reduction. To remove acoustic echo signal sufficiently, residual echo reduction is done in a multi-channel spatial filtering way. Two types of parameter optimization algorithms have been proposed based on VCD

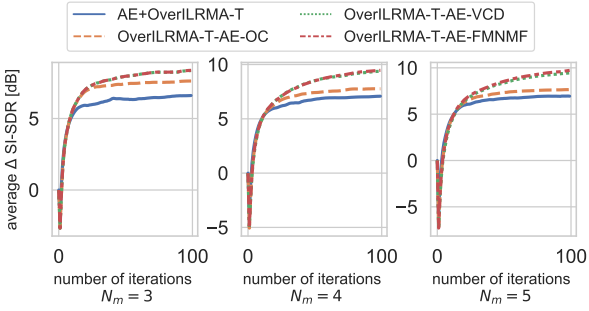

Fig. 4. Convergence speed when echo path is time-varying

and FMNMF. Experimental results showed that the proposed approach outperformed a cascade-connection based method. When acoustic impulse responses between a loudspeaker and microphones are time-varying, it is shown that residual echo reduction is effective. It is also shown that FMNMF achieved the equivalent performance with VCD even though the computational cost of the FMNMF is much smaller than that of the VCD.

REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[2] S. Makino, *Audio Source Separation*. Springer Publishing Company, Incorporated, 2018.

[3] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation*. Springer Publishing Company, Incorporated, 2007.

[4] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. SIP*, vol. 8, 2019.

[5] P. Naylor and N. Gaubitch, *Speech Dereverberation*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sept 2010.

[7] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1997, pp. 219–222 vol.1.

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

[9] M. Togami and K. Hori, "Multichannel semi-blind source separation via local gaussian modeling for acoustic echo reduction," in *EUSIPCO 2011*, Aug 2011, pp. 496–500.

[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, July 2013.

[11] M. Togami and Y. Kawaguchi, "Simultaneous optimization of acoustic echo reduction, speech dereverberation, and noise reduction against mutual interference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1612–1623, Nov 2014.

[12] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2011, pp. 189–192.

[13] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, Sept 2016.

[14] H. Kagami, H. Kameoka, and M. Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel nonnegative matrix factorization," in *ICASSP 2018*, April 2018, pp. 31–35.

[15] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "A unifying framework for blind source separation based on a joint diagonalizability constraint," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[16] ——, "Independent low-rank matrix analysis with decorrelation learning," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 288–292.

[17] M. Togami, "Multi-channel speech source separation and dereverberation with sequential integration of determined and underdetermined models," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 231–235.

[18] R. Ikeshita and T. Nakatani, "Independent vector extraction," in *2020 Acoustic Society of Japan Spring Meeting*, 2020, in Japanese.

[19] M. Togami and R. Scheibler, "Over-determined speech source separation and dereverberation," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 705–710.

[20] R. Scheibler and N. Ono, "Independent vector analysis with more microphones than sources," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 185–189.

[21] ——, "MM algorithms for joint independent subspace analysis with application to blind single and multi-source extraction," *arXiv:2004.03926*, 2020.

[22] R. Ikeshita, T. Nakatani, and S. Araki, "Overdetermined independent vector analysis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 591–595.

[23] Y. Mitsui, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Vectorwise coordinate descent algorithm for spatially regularized independent low-rank matrix analysis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 746–750.

[24] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[25] 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 Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[26] D. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004. [Online]. Available: https://doi.org/10.1198/0003130042836

[27] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 971–982, May 2013.

[28] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, "Infinite positive semidefinite tensor factorization for source separation of mixture signals," *30th International Conference on Machine Learning, ICML 2013*, pp. 1613–1621, 01 2013.

[29] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *ICASSP*, 2018, pp. 351–355.

[30] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 504–511.