Over-determined Speech Source Separation and Dereverberation

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

***Abstract*—In this paper, we propose a joint speech source separation and dereverberation technique which works well when the number of microphones is more than the number of speech sources. Microphones that exceed the number of sound sources are utilized for background noise reduction. The proposed method extends the recently proposed ILRMA-T into an over-determined technique. We reveal that an orthogonal constraint enables efficient update of a noise reduction filter in the proposed framework similar to the previously proposed over-determined speech source separation case. Secondly, the proposed method utilizes a joint diagonalization framework to reduce the residual noise signal in the output separated signal. Experimental results show that the proposed method efficiently separates speech sources in reverberant and noisy environments.**

Index Terms:dereverberation, blind speech source separation, noise reduction, joint diagonalization

I. Introduction

Signals recorded by microphones are contaminated by background noise, reverberation, and interferences. Blind speech source separation [1]–[3], e.g., independent component analysis (ICA) [4], independent vector analysis (IVA) [5], [6], and dereverberation [7] techniques, e.g., Weighted Prediction Error (WPE) [8], are required to remove unwanted signals and to recover speech quality in speech communication systems, speech diarization systems, and automatic speech recognition (ASR) systems. Although speech source separation techniques and dereverberation techniques have evolved separately, joint optimization of several techniques has not matured yet. Thus, one important challenge is how to optimize speech source separation, dereverberation, and noise reduction jointly.

Recently, joint optimization of speech separation and dereverberation has been studied [9]–[14]. These techniques can be divided into determined techniques [9], [11]–[13] and under-determined techniques [10], [14]. Approaches based on determined models assume that the number of the speech sources is equal to the number of the microphones. These approaches combine WPE [8] and a blind speech source separation such as Independent Low-Rank Matrix Analysis (ILRMA) [15], [16] jointly. All parameters can be estimated to increase the likelihood function monotonically. ILRMA-T [12], [13] optimizes a separation filter and a dereverberation filter jointly by an extension of the iterative projection (IP) method [17]. However, conventional approaches do not perform noise reduction jointly with speech source separation and dereverberation.

Another joint optimization of speech source separation and dereverberation is based on the under-determined model which assumes that the number of the speech sources is more than the number of the microphones. Because we can regard the noise signal as an additional source in the under-determined model, the noise signal can be also reduced in the speech source separation framework [10], [18]. However, it is difficult to optimize all of the parameters stably in the under-determined model such as local Gaussian model (LGM) [19], [20], because the number of parameters is larger than in the determined model. A previously proposed approach utilizes a determined model based speech source separation followed by an under-determined model based approach to increase stability of the parameter estimation in the under-determined model [14]. Because it is not efficient to estimate noise information from scratch in the under-determined model, it is required to estimate noise information also in the determined model.

Recently, joint optimization of speech source separation and noise reduction has been studied in the over-determined speech source separation context [21]–[26]. In these approaches, it is assumed that there are more microphones than the number of speech sources. In [27]–[29], it is considered to reduce the number of the microphones to the number of the speech sources. However, these methods risk removing speech sources [25]. In [24], [26], over-determined approaches based on the determined model have been proposed. A background noise signal is interpreted as one source signal, and microphones that exceed the number of speech sources are effectively utilized for reduction of the background noise signal. A computationally efficient method is also proposed based on an orthogonal constraint for the update of the separation filters of the background noise signals [24], [26], which are special cases of Independent Subspace Analysis (ISA) [30]–[32]. Because conventional over-determined based approaches do not perform speech dereverberation, one next goal is the integration of speech dereverberation in the over-determined framework.

In this paper, we propose an over-determined based joint optimization for speech source separation, dereverberation, and noise reduction. The proposed method can be regarded as an extension of ILRMA-T into an over-determined approach. Thus, we call the proposed method OverILRMA-T. We reveal that an orthogonal constraint enables efficient update of the noise reduction filter in the proposed framework similar to the over-determined speech source separation case [24], [26]. For additional noise reduction, we also propose a sequential integration of the proposed OverILRMA-T and a joint diagonalization based parameter update [33]. We conduct experiments and confirm effectiveness of the proposed method in reverberant and noisy environments with multiple speech sources.

II. Relation to prior work

As a result, the parameter optimization scheme of the proposed OverILRMA-T is identical to that of the IP-1 algorithm proposed by [34] (Ikeshita-IP-1). Ikeshita-IP-1 has been derived under the approximation that the covariance matrix of the background noise signal is an identity matrix. Due to this approximation, it is not assured that the original likelihood function increases monotonically in Ikeshita-IP-1 and the original likelihood function cannot be obtained in Ikeshita-IP-1. On contrary to Ikeshita-IP-1, the proposed method does not utilize the above assumption. It is assured that the proposed method increases the original likelihood function monotonically. In the proposed method, the original likelihood function can be calculated easily.

III. Signal model

A. Microphone input signal

The microphone input signal is defined as a convolutive mixture in the time-frequency (T-F) domain.

$$\boldsymbol{x}_{lk} = \sum_{i=1}^{N_s} \sum_{d=0}^{L_d-1} s_{i,l-d,k} \boldsymbol{a}_{idk} + \boldsymbol{n}_{lk}, \tag{1}$$

where $\boldsymbol{x}_{lk} \in \mathbb{C}^{N_m}$ (N_m is the number of the microphones), l is the frame index, k is the frequency index, L_d is the number of the tap-length of each impulse response in the T-F domain, s_{ilk} is the i-th speech source signal, $\boldsymbol{n}_{lk}$ is the multi-channel noise signal, $\boldsymbol{a}_{idk}$ is the d-th tap of the impulse response of the i-th speech source, and N_s is the number of speech sources. We assume that N_m is larger than N_s. The objective is to estimate $\{s_{ilk}\}$ from the microphone input signal $\{\boldsymbol{x}_{lk}\}$.

B. Probabilistic modeling

We assume that each speech source s_{ilk} belongs to a zero-mean time-varying Gaussian distribution:

$$p(s_{ilk}) = \mathcal{N}(0, v_{ilk}), \tag{2}$$

where v_{ilk} is the time-varying variance of the i-th speech source. $v_{ilk} \geq 0$ is modeled as follows:

$$v_{ilk} = \sum_{n=1}^{N_n} c_{iln} b_{ink}, \tag{3}$$

where 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 noise signal is also modeled as a zero-mean multi-variate time-invariant Gaussian distribution as follows:

$$p(\boldsymbol{n}_{lk}) = \mathcal{N}(0, \boldsymbol{V}_k). \tag{4}$$

The covariance matrix $\boldsymbol{V}_k$ is modeled as a low-rank matrix, $\boldsymbol{V}_k = \boldsymbol{Z}_k \boldsymbol{R}_k \boldsymbol{Z}_k^H$, where H is the Hermitian transpose of a matrix/vector, $\boldsymbol{Z}_k$ is a $N_m \times N_r$ ($N_r = N_m - N_s$) matrix, and $\boldsymbol{R}_k$ is a $N_r \times N_r$ matrix. Let $\tilde{\boldsymbol{x}}_{lk}$ be a $N_m L_d$-dimensional vector defined by $\begin{bmatrix} \boldsymbol{x}_{lk}^H & \cdots & \boldsymbol{x}_{l-L_d+1}^H \end{bmatrix}^H$. Thus, $\tilde{\boldsymbol{x}}_{lk}$ contains not only the current microphone input signal but also the past microphone input signal. The negative log likelihood function of the microphone input signal $\mathcal{L}$ can be calculated under the condition that the past microphone input signal is given [13] as follows:

$$\mathcal{L} = \sum_{lk} \sum_{i=1}^{N_s} \frac{|\boldsymbol{p}_{ik}^H \tilde{\boldsymbol{x}}_{lk}|^2}{v_{ilk}} + \log v_{ilk} + \left(\boldsymbol{P}_k^H \tilde{\boldsymbol{x}}_{lk}\right)^H \boldsymbol{R}_k^{-1} \boldsymbol{P}_k^H \tilde{\boldsymbol{x}}_{lk} \\ + \log|\det \boldsymbol{R}_k| - 2\log|\det \boldsymbol{W}_k| + \text{const.}, \tag{5}$$

where $\boldsymbol{p}_{ik}$ is the $N_m L_d$-dimensional vector which separates the i-th speech source, $\boldsymbol{P}_k$ is a $N_m L_d \times N_r$ matrix which separates the noise signal, $\boldsymbol{W}_k$ is a $N_m \times N_m$ matrix which is the upper-left partial matrix of $\mathcal{P}_k$. $\mathcal{P}_k$ is a $N_m \times N_m L_d$ matrix which is defined as $\mathcal{P}_k = \begin{pmatrix} \boldsymbol{p}_{1k} & \cdots & \boldsymbol{p}_{N_s k} & \boldsymbol{P}_k \end{pmatrix}^H$. In the proposed method, parameters $\{\mathcal{P}_k\}$, $\{\boldsymbol{R}_k\}$, $\{c_{iln}\}$, $\{b_{ink}\}$ are updated to minimize the negative log likelihood function $\mathcal{L}$.

IV. Proposed method

The proposed method updates all the parameters in an iterative way. c_{iln} and b_{ink} are updated based on non-negative matrix factorization (NMF) [15]. $\boldsymbol{p}_{ik}$ is updated based on ILRMA-T [12], [13]. $\boldsymbol{P}_k$ is updated in a similar way to the previously proposed over-determined speech source separation [24]–[26]. We reveal that an orthogonal constraint enables efficient update of the noise reduction filter $\boldsymbol{P}_k$ in the proposed framework similar to the over-determined speech source separation [24]–[26]. In [26], $\boldsymbol{R}_k$ was set to an identity matrix. We reveal that this constraint is not necessary. Rather, it is necessary to introduce $\boldsymbol{R}_k$ to ensure monotonical decrease of the cost function.

A. Speech source model estimation

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

$$b_{ink} \leftarrow b_{ink} \sqrt{\frac{\sum_t |\hat{s}_{ilk}|^2 c_{iln} \left(\sum_{n=1}^{N_n} c_{iln} b_{ink}\right)^{-2}}{\sum_t c_{iln} \left(\sum_{n=1}^{N_n} c_{iln} b_{ink}\right)^{-1}}}, \tag{6}$$

$$c_{iln} \leftarrow c_{iln} \sqrt{\frac{\sum_k |\hat{s}_{ilk}|^2 b_{ink} \left(\sum_{n=1}^{N_n} c_{iln} b_{ink}\right)^{-2}}{\sum_k b_{ink} \left(\sum_{n=1}^{N_n} c_{iln} b_{ink}\right)^{-1}}}, \tag{7}$$

where $\hat{s}_{ilk}$ is the separated signal defined as follows:

$$\hat{s}_{ilk} = \boldsymbol{p}_{ik}^H \tilde{\boldsymbol{x}}_{lk}. \tag{8}$$

B. Separation filter update based on ILRMA-T [12], [13]

The separation filter for the i-th speech signal, $\boldsymbol{p}_{ik}$, is updated as follows:

$$\boldsymbol{p}_{ik} \leftarrow \frac{\boldsymbol{Q}_{ik}^{-1}\boldsymbol{a}_{ik}}{\sqrt{\boldsymbol{a}_{ik}^{H}\boldsymbol{Q}_{ik}^{-1}\boldsymbol{a}_{ik}}}, \tag{9}$$

where

$$\boldsymbol{Q}_{ik} = \frac{1}{L}\sum_{l=1}^{L}\frac{\tilde{\boldsymbol{x}}_{lk}\tilde{\boldsymbol{x}}_{lk}^{H}}{v_{ilk}}, \tag{10}$$

$$\boldsymbol{a}_{i} = \begin{pmatrix} \overline{\boldsymbol{W}}_{k}^{-1}\boldsymbol{e}_{i} \\ \boldsymbol{0} \end{pmatrix}, \tag{11}$$

and $\boldsymbol{e}_i$ is a N_m dimensional vector in which only the m-th element takes 1 and the other elements take 0.

C. Noise separation filter update with orthogonal constraint

The derivative of the negative log-likelihood function $\mathcal{L}$ w.r.t. the noise reduction filter $\boldsymbol{P}_k$ can be calculated as follows:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{P}_k^{*}} = \boldsymbol{Q}_{nk}\boldsymbol{P}_k\boldsymbol{R}_k^{-1} - \begin{pmatrix} \boldsymbol{W}_k^{-1}\boldsymbol{E}_n \\ \boldsymbol{0}_{N_m(L_d-1)\times N_r} \end{pmatrix}, \tag{12}$$

where

$$\boldsymbol{Q}_{nk} = \frac{1}{L}\sum_{l=1}^{L}\tilde{\boldsymbol{x}}_{lk}\tilde{\boldsymbol{x}}_{lk}^{H}, \tag{13}$$

and $\boldsymbol{E}_n$ is a $N_m \times N_r$ matrix which is defined as $[\ e_{N_s} \ \ \cdots \ \ e_{N_m}\ ]$. We can obtain the following equation by taking $\frac{\partial \mathcal{L}}{\partial \boldsymbol{P}_k^{*}} = \boldsymbol{0}$ as follows:

$$\begin{pmatrix} \boldsymbol{W}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{N_m(L_d-1)\times N_m(L_d-1)} \end{pmatrix} \boldsymbol{Q}_{nk}\boldsymbol{P}_k\boldsymbol{R}_k^{-1} = \begin{pmatrix} \boldsymbol{E}_n \\ \boldsymbol{0} \end{pmatrix}, \tag{14}$$

where $\boldsymbol{I}$ is the identity matrix. Let $\overline{\boldsymbol{W}}_k$ be a $N_m \times N_m$ matrix which is equal to $\boldsymbol{W}_k$ just before $\boldsymbol{P}_k$ is updated. Because the first N_s row vectors in $\boldsymbol{W}_k$ are the same as those in $\overline{\boldsymbol{W}}_k$, each column vector in the matrix $\boldsymbol{Q}_{nk}\boldsymbol{P}_k\boldsymbol{R}_k^{-1}$ is in a linear subspace spanned by the last N_r row vectors of the matrix $\begin{pmatrix} \overline{\boldsymbol{W}}_k & \boldsymbol{0}_{N_m\times N_m(L_d-1)} \end{pmatrix}$. Thus, we can obtain the following equation:

$$\begin{pmatrix} \overline{\boldsymbol{w}}_{i,k}^{H} & \boldsymbol{0} \end{pmatrix} \left(\boldsymbol{Q}_{nk}\boldsymbol{P}_k\boldsymbol{R}_k^{-1}\right)_j = \begin{cases} 0 \text{ if } i \le N_s \\ b_{ij} \text{ otherwise}, \end{cases} \tag{15}$$

where $\overline{\boldsymbol{w}}_{i,k}^{H}$ is the i-th row vector of $\overline{\boldsymbol{W}}_k$ and $\left(\boldsymbol{Q}_{nk}\boldsymbol{P}_k\boldsymbol{R}_k^{-1}\right)_j$ is the j-th column vector of $\boldsymbol{Q}_{nk}\boldsymbol{P}_k\boldsymbol{R}_k^{-1}$. It can be summarized as follows:

$$\begin{pmatrix} \overline{\boldsymbol{W}}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \boldsymbol{Q}_{nk}\boldsymbol{P}_k\boldsymbol{R}_k^{-1} = \begin{pmatrix} \boldsymbol{E}_n \\ \boldsymbol{0} \end{pmatrix} \boldsymbol{B}_k, \tag{16}$$

where $\boldsymbol{B}_k = \{b_{ij}\}$ is a $N_r \times N_r$ matrix. $\boldsymbol{P}_k$ can be obtained as follows:

$$\boldsymbol{P}_k \leftarrow \boldsymbol{Q}_{nk}^{-1}\boldsymbol{A}_k\boldsymbol{B}_k\boldsymbol{R}_k, \tag{17}$$

where $\boldsymbol{A}_k$ is defined as follows:

$$\boldsymbol{A}_k = \begin{pmatrix} \overline{\boldsymbol{W}}_k^{-1}\boldsymbol{E}_n \\ \boldsymbol{0} \end{pmatrix}. \tag{18}$$

The remaining unknown variable is $\boldsymbol{B}_k$. $\boldsymbol{B}_k$ is obtained to fulfill (14) as follows:

$$\boldsymbol{R}_k^{H}\boldsymbol{B}_k^{H}\boldsymbol{A}_k^{H}\boldsymbol{Q}_{nk}^{-H}\boldsymbol{A}_k\boldsymbol{B}_k = \boldsymbol{I}. \tag{19}$$

Thus, $\boldsymbol{B}_k$ can be obtained as follows:

$$\boldsymbol{B}_k = \boldsymbol{F}_k^{-H}\boldsymbol{U}_k\boldsymbol{G}_k^{-1}, \tag{20}$$

where

$$\boldsymbol{G}_k\boldsymbol{G}_k^{H} = \boldsymbol{R}_k, \tag{21}$$

$$\boldsymbol{F}_k\boldsymbol{F}^{H} = \boldsymbol{A}_k^{H}\boldsymbol{Q}_{nk}^{-H}\boldsymbol{A}_k, \tag{22}$$

and $\boldsymbol{U}_k$ is a $N_r \times N_r$ arbitrary unitary matrix. The cost function $\mathcal{L}$ is not affected by the selection of $\boldsymbol{U}_k$. Thus, we set $\boldsymbol{U}_k$ to an identity matrix. Finally, we can update $\boldsymbol{P}_k$ as follows:

$$\boldsymbol{P}_k \leftarrow \boldsymbol{Q}_{nk}^{-1}\boldsymbol{A}_k\boldsymbol{F}_k^{-H}\boldsymbol{G}_k^{H}. \tag{23}$$

After updating $\boldsymbol{P}_k$, we can update $\boldsymbol{R}_k$ to minimize the cost function $\mathcal{L}$ as follows:

$$\boldsymbol{R}_k \leftarrow \frac{1}{L}\sum_{l}\boldsymbol{P}_k^{H}\tilde{\boldsymbol{x}}_{lk}\tilde{\boldsymbol{x}}_{lk}^{H}\boldsymbol{P}_k, \tag{24}$$

Proposition 1: The cost function is invariant by replacing $\boldsymbol{P}_k$ and $\boldsymbol{R}_k$ with an arbitrary non-singular square matrix $\boldsymbol{D}_k$ as follows:

$$\boldsymbol{P}_k \leftarrow \boldsymbol{P}_k\boldsymbol{D}_k, \tag{25}$$

$$\boldsymbol{R}_k \leftarrow \boldsymbol{D}_k^{H}\boldsymbol{R}_k\boldsymbol{D}_k. \tag{26}$$

Proof: In (5), the third term, the fourth term, and the fifth term depend on $\boldsymbol{R}_k$ and $\boldsymbol{P}_k$. The third term is invariant by the replacement as follows:

$$\begin{aligned} &\left(\left(\boldsymbol{P}_k\boldsymbol{D}_k\right)^{H}\tilde{\boldsymbol{x}}_{lk}\right)^{H}\left(\boldsymbol{D}_k^{H}\boldsymbol{R}_k\boldsymbol{D}_k\right)^{-1}\left(\boldsymbol{P}_k\boldsymbol{D}_k\right)^{H}\tilde{\boldsymbol{x}}_{lk} \\ &= \left(\boldsymbol{P}_k^{H}\tilde{\boldsymbol{x}}_{lk}\right)^{H}\boldsymbol{R}_k^{-1}\boldsymbol{P}_k^{H}\tilde{\boldsymbol{x}}_{lk}. \end{aligned} \tag{27}$$

The fourth term becomes

$$\log|\det\left(\boldsymbol{D}_k^{H}\boldsymbol{R}_k\boldsymbol{D}_k\right)| = \log|\det \boldsymbol{R}_k| + 2\log|\det \boldsymbol{D}_k|. \tag{28}$$

Eq. (25) does not change the N_s row vectors of $\boldsymbol{W}_k$, and Eq. (25) changes only the last N_r rows vectors. Therefore The fifth term is

$$\begin{aligned} &-2\log\left|\det\begin{pmatrix} \boldsymbol{I}_{N_s\times N_s} & \boldsymbol{0} \\ 0 & \boldsymbol{D}_k \end{pmatrix}\boldsymbol{W}_k\right| \\ &= -2\log|\det \boldsymbol{W}_k| - 2\log|\det \boldsymbol{D}_k|. \end{aligned} \tag{29}$$

Summation of the fourth term and the fifth term is invariant by the replacement. Thus, $\mathcal{L}$ defined by (5) is invariant by the replacement. ■

Based on the proposition 1, we can obtain the following simpler update of $\boldsymbol{P}_k$ by setting $\boldsymbol{D}_k$ to $\boldsymbol{G}_k^{-H}\boldsymbol{F}_k^{H}$:

$$\boldsymbol{P}_k \leftarrow \boldsymbol{P}_k\boldsymbol{D}_k = \boldsymbol{Q}_{nk}^{-1}\begin{pmatrix} \overline{\boldsymbol{W}}_k^{-1}\boldsymbol{E}_n \\ \boldsymbol{0} \end{pmatrix}. \tag{30}$$

Furthermore, by setting $\boldsymbol{D}_k$ to $\boldsymbol{G}_k^{-H}\boldsymbol{F}_k^H\boldsymbol{W}_{nk}^{-H}$ ($\boldsymbol{W}_{nk}$ is the left-bottom $N_r \times N_r$ matrix of $\boldsymbol{W}_k$), we can update $\boldsymbol{P}_k$ more efficiently as follows:

$$\boldsymbol{P}_k = \begin{pmatrix} \boldsymbol{A}_{nk} \\ -\boldsymbol{I} \\ \boldsymbol{J}_{nk,3}\boldsymbol{J}_{nk,1}^{-1}\boldsymbol{A}_{nk} - \boldsymbol{J}_{nk,3}\boldsymbol{J}_{nk,1}^{-1}\boldsymbol{E}_n \end{pmatrix}, \tag{31}$$

where

$$\begin{pmatrix} \boldsymbol{J}_{nk,1,N_m\times N_m} & \boldsymbol{J}_{nk,2} \\ \boldsymbol{J}_{nk,3} & \boldsymbol{J}_{nk,4} \end{pmatrix} = \boldsymbol{Q}_{nk}^{-1}, \tag{32}$$

$$\boldsymbol{A}_{nk} = (\overline{\boldsymbol{W}}_{s,k,N_s\times N_m}\boldsymbol{J}_{nk,1}^{-1}\boldsymbol{E}_s)^{-1}\overline{\boldsymbol{W}}_{s,k}\boldsymbol{J}_{nk,1}^{-1}\boldsymbol{E}_n. \tag{33}$$

(31) is corresponding to the extension of the orthogonal constraint for multi-channel speech source separation [24], [26] into joint speech source separation and dereverberation. In addition to the speech source separation filter, the dereverberation filter can be updated efficiently with the proposed orthogonal constraint. Even when $\boldsymbol{P}_k$ is updated with $\boldsymbol{D}_k$, $\boldsymbol{R}_k$ can be obtained based on (24). Because $\boldsymbol{R}_k$ does not affect updates of $\boldsymbol{P}_k$ based on (31) and $\boldsymbol{p}_{ik}$ based on (9), we can actually skip update of $\boldsymbol{R}_k$. After the iterative parameter update, the output signal $\boldsymbol{y}_{ilk} \in \mathbb{C}^{N_m}$ is obtained via projection-back as follows:

$$\boldsymbol{y}_{ilk} = \left(\boldsymbol{W}_k^{-1}\right)_i \hat{s}_{ilk}, \tag{34}$$

where $\left(\boldsymbol{W}_k^{-1}\right)_i$ is the i-th column vector of $\boldsymbol{W}_k^{-1}$. We call this algorithm **OverILRMA-T-1**. The algorithm of **OverILRMA-T-1** is summarized in Algorithm 1.

Algorithm 1 Algorithm of **OverILRMA-T-1**

```
Input: $\{\tilde{\boldsymbol{x}}_{lk}\}$, Initialized value of $\{\hat{s}_{ilk}\}$, $\{\boldsymbol{P}_k\}$, $\{c_{iln}\}$, and $\{b_{ink}\}$
Output: Separated signal $\boldsymbol{y}_{ilk} \in \mathbb{C}^{N_m}$
$\begin{pmatrix} \boldsymbol{J}_{nk,1} & \boldsymbol{J}_{nk,2} \\ \boldsymbol{J}_{nk,3} & \boldsymbol{J}_{nk,4} \end{pmatrix} = \boldsymbol{Q}_{nk}^{-1}$, with $\boldsymbol{Q}_{nk} = \frac{1}{L}\sum_{l=1}^{L}\tilde{\boldsymbol{x}}_{lk}\tilde{\boldsymbol{x}}_{lk}^H$, $\forall k$
for $t = 1$ to $N_t$ do
    for $i = 1$ to $N_s$ do
        Update NMF parameters $b_{ink}$ and $c_{iln}$, $\forall n, \forall k$ based on (6) and (7).
        Update $\boldsymbol{p}_{ik}$, $\forall k$ based on (9).
    end for
    Update $\boldsymbol{P}_k$, $\forall k$ based on (31).
    for $i = 1$ to $N_s$ do
        Estimate $\hat{s}_{ilk}$ $\forall l, \forall k$ based on (8).
    end for
end for
Perform projection back based on (34).
```

D. Additional noise reduction based on joint diagonalization

Typically, it is not sufficient to reduce noise based on a determined model. We adopt additional removal of residual noise signal by utilizing a joint diagonalization (JD) technique [33]. After estimating parameters based on the proposed OverILRMA-T-1, a new parameter $\boldsymbol{r}_k$ is updated. The cost function of the proposed OverILRMA-T with JD $\mathcal{L}_{JD}$ (**OverILRMA-T-2**) is formulated as follows:

$$\begin{aligned} \mathcal{L}_{JD} = & \sum_{lk}\sum_{i=1}^{N_m}\frac{|\boldsymbol{p}_{ik}^H\tilde{\boldsymbol{x}}_{lk}|^2}{\left(\sum_{s=1}^{N_s} v_{ilk}r_{isk} + \sum_{s=N_s+1}^{N_m} r_{isk}\right)} \\ & + \log\left(\sum_{s=1}^{N_s} v_{ilk}r_{isk} + \sum_{s=N_s+1}^{N_m} r_{isk}\right) \\ & - 2\log|\det \boldsymbol{W}_k| + \text{const.}, \end{aligned} \tag{35}$$

$\boldsymbol{r}_{isk}$ is updated in an iterative manner to increase $\mathcal{L}_{JD}$ monotonically based on joint diagonalization [33]. We also expect that it is better to set $r_{isk} = 0$ when $s \leq N_s$ and $i \neq s$, because we can prevent remixing from the s ($\neq i$)-th speech stream to the i-th speech stream when $r_{isk} = 0$. In this case, we can obtain the following simplified cost function:

$$\begin{aligned} \mathcal{L}_{JD2} = & \sum_{lk}\sum_{i=1}^{N_s}\frac{|\boldsymbol{p}_{ik}^H\tilde{\boldsymbol{x}}_{lk}|^2}{v_{ilk} + r_{ik}} + \log\left(v_{ilk} + r_{ik}\right) - 2\log|\det \boldsymbol{W}_k| \\ & + \left(\boldsymbol{P}_{nk}^H\tilde{\boldsymbol{x}}_{lk}\right)^H \boldsymbol{R}_{nk}^{-1}\boldsymbol{P}_{nk}^H\tilde{\boldsymbol{x}}_{lk} + \log|\det \boldsymbol{R}_{nk}|, \end{aligned} \tag{36}$$

r_{ik} stands for the amount of the noise signal at the i-th output signal. $\{r_{ik}\}$, $\{c_{iln}\}$, and $\{b_{ink}\}$ are updated as follows:

$$r_{ik} \leftarrow r_{ik}\sqrt{\frac{\sum_t |\hat{s}_{ilk}|^2\left(\sum_{n=1}^{N_n} c_{iln}b_{ink} + r_{ik}\right)^{-2}}{\sum_t\left(\sum_{n=1}^{N_n} c_{iln}b_{ink} + r_{ik}\right)^{-1}}}, \tag{37}$$

$$b_{ink} \leftarrow b_{ink}\sqrt{\frac{\sum_t |\hat{s}_{ilk}|^2 c_{iln}\left(\sum_{n=1}^{N_n} c_{iln}b_{ink} + r_{ik}\right)^{-2}}{\sum_t c_{iln}\left(\sum_{n=1}^{N_n} c_{iln}b_{ink} + r_{ik}\right)^{-1}}}, \tag{38}$$

$$c_{iln} \leftarrow c_{iln}\sqrt{\frac{\sum_k |\hat{s}_{ilk}|^2 b_{ink}\left(\sum_{n=1}^{N_n} c_{iln}b_{ink} + r_{ik}\right)^{-2}}{\sum_k b_{ink}\left(\sum_{n=1}^{N_n} c_{iln}b_{ink} + r_{ik}\right)^{-1}}}. \tag{39}$$

The other parameters, i.e., $\boldsymbol{P}_k$ and $\boldsymbol{p}_{ik}$, are updated in the same way as OverILRMA-T-1 based on (31) and (9). The output signal can be obtained as follows:

$$\boldsymbol{y}_{ilk} = \frac{v_{ilk}}{v_{ilk} + r_{ik}}\left(\boldsymbol{W}_k^{-1}\right)_i \boldsymbol{p}_{ik}^H\tilde{\boldsymbol{x}}_{lk}. \tag{40}$$

We call this algorithm **OverILRMA-T-3**.

TABLE I
SIMULATION CONFIGURATIONS

T_{60} [sec]	Max order	Absorption
0.33	17	0.35
0.70	30	0.19

TABLE II
EVALUATION RESULTS WHEN $T_{60} = 0.33$ [SEC]

Approach	Δ SDR (dB)	Δ SIR (dB)	Δ CD (dB)	Δ LLR	Δ FwSegSNR (dB)	Δ SRMR
ILRMA-T	4.63	9.50	-0.36	-0.12	0.87	1.36
PCA+ILRMA-T	4.60	8.86	-0.33	-0.10	0.94	1.29
OverILRMA-T-1	5.36	9.63	-0.39	-0.13	1.21	1.64
OverILRMA-T-2	4.69	7.49	-0.46	-0.16	0.30	2.24
OverILRMA-T-3	**5.59**	**10.19**	**-0.83**	**-0.23**	**1.55**	**2.64**

TABLE III
EVALUATION RESULTS WHEN $T_{60} = 0.70$ [SEC]

Approach	Δ SDR (dB)	Δ SIR (dB)	Δ CD (dB)	Δ LLR	Δ FwSegSNR (dB)	Δ SRMR
ILRMA-T	4.63	8.68	-0.37	-0.11	0.58	1.45
PCA+ILRMA-T	4.05	7.31	-0.32	-0.08	0.30	1.23
OverILRMA-T-1	5.47	9.29	-0.42	-0.12	1.03	1.79
OverILRMA-T-2	4.92	7.30	-0.51	-0.15	0.34	2.41
OverILRMA-T-3	**5.77**	**9.80**	**-0.87**	**-0.23**	**1.56**	**2.64**

V. EXPERIMENT

A. Setup

Performances of speech source separation, dereverberation, and noise reduction were evaluated with simulated data made by `Pyroomacoustics` [35]. Anechoic speech sources were extracted from the CMU Sphinx database [36]. `Pyroomacoustics` simulated reverberant mixtures in a $10 \times 10 \times 10$ m room with two configurations shown in Table I. Sampling rate was 16000 Hz. Signal to Noise Ratio (SNR) between speech sources and background noise was set to 10 dB or 20 dB. The number of the microphones N_m was 4. We used a square microphone array with a side of $4\sqrt{2}$ cm. The number of the speech sources N_s was set to 2. Distance between microphones and talkers was set to 1 m. 100 reverberant mixtures were simulated for each condition. The azimuth of each speech source was randomly selected so that the azimuth difference between speech sources is more than 30 degree. The utterances and the talkers were randomly selected for each mixture. Frame size was 1024. Frame shift was 512. L_d was set to 4. The number of the parameter updates N_t was set to 40. In OverILRMA-T-2 and OverILRMA-T-3, the parameters were updated in the same way as OverILRMA-T-1 in the first 30 iterations. After that, the parameters were updated to minimize the cost functions with joint diagonalization, i.e., (35) and (36) by introducing $\boldsymbol{r}$.

B. Results

We utilized the signal-to-distortion ratio (SDR) (dB), the signal-to-interference ratio (SIR) (dB), the cepstrum distance (CD) (dB), the log likelihood ratio (LLR), the frequency-weighted segmental SNR (FWSeg.SNR) (dB), and the speech-to-reverberation modulation energy ratio (SRMR) as evaluation measures. SDR and SIR were calculated by BSS_EVAL [37]. The other measures were calculated in the same way as [38]. Evaluation results are shown in Table II and Table III. The proposed methods were compared with ILRMA-T [13]. After ILRMA-T, the N_s strongest outputs were selected similarly to [24]. Additionally, we also compared principal component analysis (PCA) + ILRMA-T which reduces the number of the input microphones from N_m to N_s similarly to [28], [29]. It is shown that the proposed OverILRMA-T-3 outperformed the other methods. From comparison of OverILRMA-T-1 with ILRMA-T and PCA+ILRMA-T, it is shown that the proposed background noise reduction is more effective than the original ILRMA-T. From comparison between OverILRMA-T-1 and OverILRMA-T-3, it is shown that residual noise reduction with additional parameter update with the joint diagonalization is effective. On the other hand, by comparison of the results of OverILRMA-T-2 and OverILRMA-T-3, it is shown that it is effective to utilize joint diagonalization only for residual noise reduction.

VI. CONCLUSIONS

In this paper, we proposed a joint optimization technique of speech source separation, dereverberation, and noise reduction in which it is assumed that the number of the microphones is more than the number of the speech sources. We revealed that it is possible to update the noise separation filter efficiently with monotonical decrease of the cost function based on the proposed orthogonal constraint similarly to the conventional over-determined speech source separation. Additionally, we proposed residual noise reduction based on joint diagonalization. Experimental results showed that the proposed method outperformed ILRMA-T based joint speech source separation and dereverberation with a determined model. It is also shown that joint diagonalization is effective when it is utilized for only residual noise reduction.

REFERENCES

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

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

[3] 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.

[4] P. Common, "Independent component analysis, a new concept ?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, April 1994.

[5] A. Hiroe, "Solution of permutation problem in frequency domain ica using multivariate probability density functions," in *Proceedings ICA*, Mar. 2006, pp. 601–608.

[6] T. Kim, H. Attias, S.-Y. Lee, and T.-W. Lee, "Independent vector analysis: an extension of ica to multivariate components," in *Proceedings ICA*, Mar. 2006, pp. 165–172.

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

[8] 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.

[9] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 69–84, Jan 2011.

[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] 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.

[12] 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.

[13] ——, "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.

[14] 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.

[15] 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.

[16] ——, *Determined Blind Source separation with Independent Low-Rank Matrix Analysis*. Springer Publishing Company, Incorporated, 2018, ch. 6, pp. 125–155.

[17] 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.

[18] M. Togami, "Multi-channel time-varying covariance matrix model for late reverberation reduction," *arXiv:1910.08710*, 2019.

[19] 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.

[20] 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.

[21] Y. Kubo, N. Takamune, D. Kitamura, and H. Saruwatari, "Efficient full-rank spatial covariance estimation using independent low-rank matrix analysis for blind source separation," in *Proc. EUSIPCO 2019*, Sep 2019, pp. 1814–1818.

[22] Z. Koldovský, P. Tichavský, and V. Kautský, "Orthogonally constrained independent component extraction: Blind mpdr beamforming," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 1155–1159.

[23] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1050–1064, 2019.

[24] 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.

[25] R. Scheibler and N. Ono, "MM algorithms for joint independent subspace analysis with application to blind single and multi-source extraction," *arXiv:2004.03926*, 2020.

[26] 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.

[27] T. Nishikawa, H. Abe, H. Saruwatari, and K. Shikano, "Overdetermined blind separation for convolutive mixtures of speech based on multistage ica using subarray processing," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–225.

[28] C. Osterwise and S. L. Grant, "On over-determined frequency domain bss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 5, pp. 956–966, 2014.

[29] M. Joho, H. Mathis, and R. Lambert, "Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture," 08 2000.

[30] R. F. Silva, S. M. Plis, T. Adali, and V. D. Calhoun, "Multidataset independent subspace analysis extends independent vector analysis," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 2864–2868.

[31] D. Lahat and C. Jutten, "Joint independent subspace analysis using second-order statistics," *IEEE Transactions on Signal Processing*, vol. 64, no. 18, pp. 4891–4904, 2016.

[32] C. J. D.Lahat, "Joint independent subspace analysis: A quasi-newton algorithm," in *Proc. LVA/ICA*, vol. 9237, no. 1, 2015, pp. 111–118.

[33] 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.

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

[35] 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.

[36] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *in 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[37] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[38] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *Eurasip Journal on Advances in Signal Processing*, vol. 2016, pp. 1–19, 2016.