# Robust Demixing Filter Update Algorithm Based on Microphone-wise Coordinate Descent for Independent Deeply Learned Matrix Analysis

Naoki Makishima*, Norihiro Takamune*, Daichi Kitamura†, Hiroshi Saruwatari*, Yu Takahashi‡, and Kazunobu Kondo‡
* The University of Tokyo, Tokyo, Japan
† National Institute of Technology, Kagawa College, Kagawa, Japan
‡ Yamaha Corporation, Shizuoka, Japan

***Abstract*—In this paper, we propose a robust demixing filter update algorithm for audio source separation, which is the task of recovering source signals from multichannel mixtures observed in a microphone array. Recently, independent deeply learned matrix analysis (IDLMA) has been proposed as a state-of-the-art separation method. IDLMA utilizes the deep neural network (DNN) inference of source models and the blind estimation of demixing filters based on sources' independence. In conventional IDLMA, iterative projection (IP) is exploited to estimate the demixing filters. Although IP is a fast algorithm, when a specific source model is not accurate owing to an unfavorable SNR condition, the subsequent update of filters will fail. This is because IP updates the demixing filters in a sourcewise manner, where only one source model is used for each update. In this paper, we derive a new microphone-wise update algorithm that exploits all information of the source models simultaneously for each update. The microphone-wise update problem cannot be solved by IP, but instead, a new type of vectorwise coordinate descent algorithm is introduced into the proposed algorithm to realize convergence-guaranteed parameter estimation. Experimental results show that the proposed update algorithm achieves better separation performance than IP.**

## I. Introduction

Audio source separation aims to recover source signals from multichannel mixtures observed using a microphone array [1]. Many types of algorithm have been proposed, e.g., unsupervised (blind) methods [2]–[13] and supervised (informed) methods [14]–[17]. Independent deeply learned matrix analysis (IDLMA) [18], [19] is a state-of-the-art source separation method combining the blind estimation of a spatial model (demixing filters) and the supervised deep neural network (DNN) inference of source models. This paper also addresses the issue on improvement of IDLMA in a theoretical aspect.

In conventional separation methods including IDLMA, a computationally efficient algorithm called *iterative projection* (IP) [5], [20] is exploited to estimate the demixing filters. Although IP is a fast algorithm, when a specific source model is not accurate owing to an unfavorable SNR condition, the successive update of filters will fail hereafter. This is because IP updates the demixing filters in a sourcewise manner, where only one source model is used for each update. Therefore, development of a robust algorithm to obtain the demixing filters is a problem requiring urgent attention.

In order to resolve the above-mentioned problem, in this paper, we derive a new *microphone-wise* update algorithm that exploits all information of the source models simultaneously for each update. The microphone-wise update problem cannot be solved by IP, but instead, a new type of vectorwise coordinate descent (VCD) algorithm is introduced into the proposed algorithm to realize convergence-guaranteed parameter estimation. Experimental results show that the proposed update algorithm achieves better separation performance than IP. The main contribution of this paper is the theoretical derivation and experimental evaluation of the new microphone-wise update algorithm. Note that further application to the DNN-based automatic selection of sourcewise and microphone-wise update algorithms is beyond the scope of this paper and is discussed in [21].

## II. Conventional Method

### A. Formulation

We denote the numbers of microphones and sources as $M$ and $N$, respectively. In this paper, we assume $M = N$ for simplicity. The short-time Fourier transforms (STFTs) of the multichannel source, observed, and estimated signals are defined as

$$\boldsymbol{s}_{ij} = (s_{ij1}, \ldots, s_{ijN}) \ , \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijM}) \ , \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijN}) \ , \tag{3}$$

where $i = 1, \ldots, I; j = 1, \ldots, J; n = 1, \ldots, N$; and $m = 1, \ldots, M$ are the indexes of the frequency bins, time frames, sources, and observed microphones, respectively, and denotes the transpose. We also denote their spectrograms as $\boldsymbol{S}_n \in \mathbb{C}^{I\times J}$, $\boldsymbol{X}_n \in \mathbb{C}^{I\times J}$, and $\boldsymbol{Y}_n \in \mathbb{C}^{I\times J}$, whose elements are $s_{ijn}, x_{ijn}$, and $y_{ijn}$, respectively. When the mixing system is time-invariant and the window length in the STFT is sufficiently longer than the impulse response, the following instantaneous mixing model holds:

$$\boldsymbol{x}_{ij} = \boldsymbol{A}_i \boldsymbol{s}_{ij}, \tag{4}$$

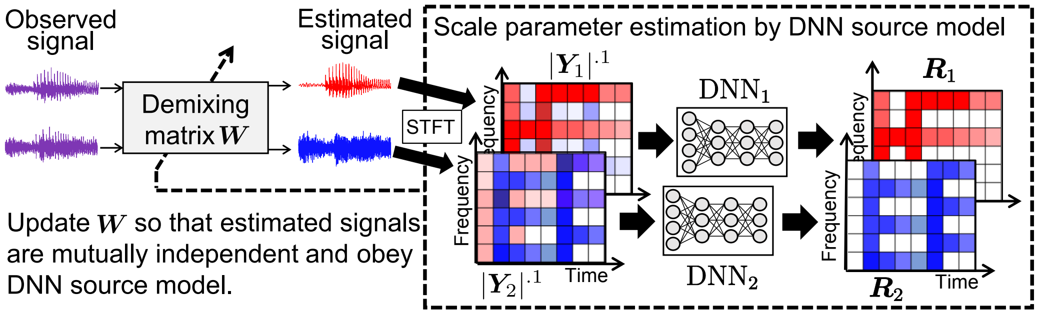


Fig. 1. Overview of IDLMA.

where $\boldsymbol{A}_i = (\boldsymbol{a}_{i1}, \ldots, \boldsymbol{a}_{iN}) \in \mathbb{C}^{I\times J}$ is the mixing matrix and $\boldsymbol{a}_{in}$ is the steering vector of the $n$th source. When $\boldsymbol{A}_i$ is a nonsingular matrix, the demixing matrix (inverse of the mixing matrix) exists and the observed signals are separated as

$$\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}, \tag{5}$$

where $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iN})^{\mathrm{H}} \in \mathbb{C}^{I\times J} = \boldsymbol{A}_i^{-1}$ denotes the demixing matrix, $\boldsymbol{w}_{in}^{\mathrm{H}}$ denotes the demixing filter for the $n$th source, and $^{\mathrm{H}}$ denotes the Hermitian transpose.

### B. Generative Model and Cost Function

In IDLMA, the following univariate complex Gaussian distribution is assumed as a source generative model:

$$\begin{aligned} p(\boldsymbol{Y}_n) &= \prod_{ij} p(y_{ij}) \\ &= \prod_{ij} \frac{1}{\pi {r_{ijn}}^2} \exp(-\frac{|y_{ijn}|^2}{{r_{ijn}}^2}), \end{aligned} \tag{6}$$

where $r_{ijn}$ denotes the scale parameter (source model) of the Gaussian distribution and $y_{ij}$ is mutually independent w.r.t. $i$ and $j$. We define the scale parameter matrix as $\boldsymbol{R}_n \in \mathbb{R}^{I\times J}$, whose elements are $r_{ijn}$. The marginal distribution of (6) w.r.t. $j$ is super-Gaussian when the scale parameter fluctuates and is not constant w.r.t. the time frame.

The cost function of IDLMA is the negative log-likelihood of observed signals, whose minimization is equivalent to the maximization of the independence between sources. On the basis of (6), the cost function is obtained as

$$\begin{aligned} \mathcal{L}(\boldsymbol{W}) &= -\log p(\boldsymbol{X}) \\ &= -\log p(\boldsymbol{Y}) - J \sum_i \log |\det \boldsymbol{W}_i|^2 \\ &\overset{c}{=} \sum_{i,j,n} \left[ \frac{|\boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{x}_{ij}|^2}{{r_{ijn}}^2} + 2 \log r_{ijn} \right] \\ &\quad - J \sum_i \log |\det \boldsymbol{W}_i|^2, \end{aligned} \tag{7}$$

where $\overset{c}{=}$ denotes the equality up to addition by a constant, $\boldsymbol{W} = \{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_I\}$ is the set of demixing matrices, $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M\}$ and $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N\}$ are the sets of the observed and estimated signals, respectively, and we used the variable transformation from $\boldsymbol{x}_{ij}$ to $\boldsymbol{y}_{ij}$ on the basis of (5). The aim of IDLMA is to blindly estimate $\boldsymbol{W}_i$ only from the observed mixtures with the assistance of a DNN. An overview of the separation process of IDLMA is shown in Fig. 1.

### C. Row-wise Update Rule of Demixing Matrix

In [5], [20], a fast and convergence-guaranteed algorithm called IP was proposed, which can be applied to the sum of a negative log-determinant and a quadratic form. Therefore, given the source scale parameter $r_{ijn}$, (7) is minimized by IP w.r.t. $\boldsymbol{W}_i$ and the update rule of $\boldsymbol{W}_i$ is obtained as

$$\boldsymbol{Q}_{in} = \frac{1}{J} \sum_j \frac{\boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{H}}}{{r_{ijn}}^2}, \tag{8}$$

$$\boldsymbol{w}_{in} = (\boldsymbol{W}_i \boldsymbol{Q}_{in})^{-1} \boldsymbol{e}_n, \tag{9}$$

$$\boldsymbol{w}_{in} = \frac{\boldsymbol{w}_{in}}{\sqrt{\boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{Q}_{in} \boldsymbol{w}_{in}}}, \tag{10}$$

where $\boldsymbol{e}_n$ denotes the unit vector with the $n$th element equal to unity.

### D. Update Rule of Scale Parameter Matrix by DNN

DNN$_n$ is pretrained so that the scale parameter of the source signal $\tilde{\boldsymbol{S}}_n \in \mathbb{C}^{I\times J}$ is predicted from an input spectrogram $|\boldsymbol{X}|^{.1}$, where $\boldsymbol{X} \in \mathbb{C}^{I\times J}$ is a mixture of complex-valued spectrograms in the training data and $|\cdot|^{.1}$ for matrices denotes the element-wise absolute operation. $\boldsymbol{X}$ is prepared by mixing $\tilde{\boldsymbol{S}}_n$ with a random amplitude to simulate multiple SNR conditions [19].

We denote the DNN output as $\mathrm{DNN}_n(\cdot)$. When we define the output scale parameter matrix as $\boldsymbol{D}_n = \mathrm{DNN}_n(|\tilde{\boldsymbol{X}}|^{.1}) \approx \boldsymbol{R}_n$, the loss function of DNN$_n$ is defined as

$$\mathrm{L}(\boldsymbol{D}_n) = \sum_{i,j} \frac{|\tilde{s}_{ijn}|^2 + \delta}{{d_{ijn}}^2 + \delta} - \log \frac{|\tilde{s}_{ijn}|^2 + \delta}{{d_{ijn}}^2 + \delta} - 1, \tag{11}$$

where $\tilde{s}_{ijn}$ and $d_{ijn}$ are the elements of $\tilde{\boldsymbol{S}}_n$ and $\boldsymbol{D}_n$, respectively, and $\delta$ is a small value to avoid division by zero. Since minimizing (11) corresponds to a simulation for the maximum likelihood estimation of $r_{ijn}$ in (7) (only limited to the training data), DNN$_n$ can be approximately interpreted as an appropriate source model based on (6).

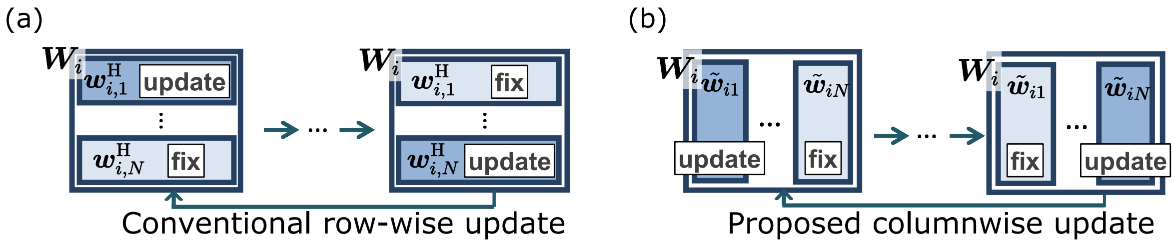


Fig. 2. (a) Conventional row-wise update of $\boldsymbol{W}_i$. (b) Proposed columnwise update of $\boldsymbol{W}_i$.

In inference for open data, the scale parameter matrix $\boldsymbol{R}_n$ is estimated by the pretrained DNN$_n$ as follows:

$$\boldsymbol{R}_n \leftarrow \mathrm{DNN_n}(|\boldsymbol{Y}_n|^{\cdot 1}), \tag{12}$$

$$r_{ijn} \leftarrow \max(r_{ijn}, \epsilon), \tag{13}$$

where $\epsilon$ is a small value to increase the numerical stability of IP. The input of the DNN, $|\boldsymbol{Y}_n|^{\cdot 1}$, is the spectrogram of each separated signal temporally obtained through the update of $\boldsymbol{W}_i$. Thus, in IDLMA, $r_{ijn}$ and $\boldsymbol{W}_i$ are alternatively updated by DNN and IP to output the most independent sources.

## III. Proposed Method

### A. Motivation

In IDLMA, the spatial model is updated by IP. Our preliminary experiments show that the separation performance of IDLMA is affected by the update order of the demixing filter in IP. This is because when the demixing filter with an inaccurate source model is updated first, the subsequent update fails. In this paper, we propose a new microphone-wise (i.e., columnwise in $\boldsymbol{W}_i$) update algorithm of the demixing matrix that simultaneously exploits all of the source models for each update. Fig. 2 illustrates the difference between the conventional IP and the proposed columnwise update algorithm. Since the cost function w.r.t. the columnvector of $\boldsymbol{W}_i$ cannot be minimized by IP, we employ the VCD algorithm [23] to derive the update rule.

### B. Cost Function w.r.t. Column Vector

We denote the column vector of $\boldsymbol{W}_i$ as $\boldsymbol{W}_i = (\tilde{\boldsymbol{w}}_{i1}, \ldots, \tilde{\boldsymbol{w}}_{iM})$, where $\tilde{\boldsymbol{w}}_{im}$ is a microphone-wise vector although $\boldsymbol{w}_{in}$ is a sourcewise (row) vector. The cost function (7) is rewritten using $\tilde{\boldsymbol{w}}_{im}$ as follows:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W})/J &\overset{c}{=} \sum_i \Big[ \sum_n \boldsymbol{w}_{in}^{\mathrm{H}} \boldsymbol{Q}_{in} \boldsymbol{w}_{in} - \log |\det \boldsymbol{W}_i|^2 \Big] \\
&= \sum_i \Bigg[ \sum_{n=1}^{N} \sum_{m_1=1}^{N} \sum_{m_2=1}^{N} w_{inm_1}^{*} Q_{inm_1m_2} w_{inm_2} \\
&\qquad - \log |\det \boldsymbol{W}_i|^2 \Bigg] \\
&= \sum_i \Bigg[ \sum_{n=1}^{N} \sum_{m=1}^{N} w_{inm}^{*} Q_{inmm} w_{inm} \\
&\qquad + \sum_{n=1}^{N} \sum_{m_1=1}^{N} \sum_{m_2 \neq m_1}^{N} w_{inm_1}^{*} Q_{inm_1m_2} w_{inm_2} \\
&\qquad - \log |\det \boldsymbol{W}_i|^2 \Bigg] \\
&\overset{c}{=} \sum_i \Big[ \tilde{\boldsymbol{w}}_{im}^{\mathrm{H}} \tilde{\boldsymbol{Q}}_{im} \tilde{\boldsymbol{w}}_{im} + \tilde{\boldsymbol{w}}_{im}^{\mathrm{H}} \tilde{\boldsymbol{h}}_{im} + \tilde{\boldsymbol{h}}_{im}^{\mathrm{H}} \tilde{\boldsymbol{w}}_{im} \\
&\qquad - \log |\det \boldsymbol{W}_i|^2 \Big] \\
&\overset{c}{=} \sum_i \Big[ \tilde{\boldsymbol{w}}_{im}^{\mathrm{H}} \tilde{\boldsymbol{Q}}_{im} \tilde{\boldsymbol{w}}_{im} + \tilde{\boldsymbol{w}}_{im}^{\mathrm{H}} \tilde{\boldsymbol{h}}_{im} + \tilde{\boldsymbol{h}}_{im}^{\mathrm{H}} \tilde{\boldsymbol{w}}_{im} \\
&\qquad - \log |\boldsymbol{b}_{im}^{\mathrm{H}} \tilde{\boldsymbol{w}}_{im}|^2 \Big],
\end{aligned} \tag{14}$$

where

$$\tilde{\boldsymbol{Q}}_{im} = \mathrm{diag}(Q_{i1mm}, \ldots, Q_{iNmm}), \tag{15}$$

$$\tilde{\boldsymbol{h}}_{im} = \Bigg( \sum_{m' \neq m} w_{i1m'}^{*} Q_{i1m'm}, \ldots, \sum_{m' \neq m} w_{iNm'}^{*} Q_{iNm'm} \Bigg). \tag{16}$$

Here, $Q_{inm_1m_2}$ is the $(m_1, m_2)$th element of $\boldsymbol{Q}_{in}$, $w_{inm}$ is the $m$th element of $\boldsymbol{w}_{in}$, $^*$ denotes the complex conjugate, and $\boldsymbol{B}_i = (\boldsymbol{b}_{i1}, \ldots, \boldsymbol{b}_{iM})^{\mathrm{H}}$ is the adjugate matrix of $\boldsymbol{W}_i$.

### C. Columnwise Update Rule of Demixing Matrix

IP cannot be applied to the minimization of the cost function (14) unlike the case of $\boldsymbol{w}_{in}$ because (14) includes the linear terms w.r.t. $\tilde{\boldsymbol{w}}_{im}$, i.e., $\tilde{\boldsymbol{w}}_{im}^{\mathrm{H}} \tilde{\boldsymbol{h}}_{im}$ and $\tilde{\boldsymbol{h}}_{im}^{\mathrm{H}} \tilde{\boldsymbol{w}}_{im}$. Accordingly, we derive a new coordinate descent algorithm for (14), where $\tilde{\boldsymbol{w}}_{im}$ for each $m$ (the microphone number) is updated by finding a stationary point of the cost function w.r.t. $\tilde{\boldsymbol{w}}_{im}$ under $\tilde{\boldsymbol{w}}_{im'}$ fixed $(m' \neq m)$. Since $\boldsymbol{b}_{im}^{\mathrm{H}}$ is independent of $\tilde{\boldsymbol{w}}_{im}$, the partial derivative of (14) w.r.t. $\tilde{\boldsymbol{w}}_{im}^{*}$ is obtained as

$$\frac{1}{J} \frac{\partial \mathcal{L}(\boldsymbol{W})}{\partial \tilde{\boldsymbol{w}}_{im}^{*}} = \tilde{\boldsymbol{Q}}_{im} \tilde{\boldsymbol{w}}_{im} + \tilde{\boldsymbol{h}}_{im} - \frac{\boldsymbol{b}_{im}}{\tilde{\boldsymbol{w}}_{im}^{\mathrm{H}} \boldsymbol{b}_{im}}. \tag{17}$$

Hereafter, we derive the update rule of $\tilde{\boldsymbol{w}}_{im}$ as in [23]. From $\partial \mathcal{L}(\boldsymbol{W}) / \partial \tilde{\boldsymbol{w}}_{im}^{*} = 0$, we obtain the following equation that satisfies the stationary-point condition:

$$\tilde{\boldsymbol{w}}_{im} = \tilde{\boldsymbol{Q}}_{im}^{-1} (\beta_{im} \boldsymbol{b}_{im} - \tilde{\boldsymbol{h}}_{im}), \tag{18}$$

where $\beta_{im} = 1/(\tilde{\boldsymbol{w}}_{im}^{\mathrm{H}} \boldsymbol{b}_{im})$. From the definition of $\beta_{im}$, we have

$$\beta_{im} \tilde{\boldsymbol{w}}_{im}^{\mathrm{H}} \boldsymbol{b}_{im} - 1 = 0. \tag{19}$$

By substituting (18) for (19), we obtain

$$\boldsymbol{b}_{im}^{\mathrm{H}} \tilde{\boldsymbol{Q}}_{im}^{-1} \boldsymbol{b}_{im} |\beta_{im}|^2 - \tilde{\boldsymbol{h}}_{im}^{\mathrm{H}} \tilde{\boldsymbol{Q}}_{im}^{-1} \boldsymbol{b}_{im} \beta_{im} - 1 = 0. \tag{20}$$

Since the first and third terms in (20) are real numbers, the second term in (20) must also be a real number, which satisfies

$$\mathrm{Im}\left[\tilde{\boldsymbol{h}}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\boldsymbol{b}_{im}\beta_{im}\right]=0, \tag{21}$$

where $\mathrm{Im}[\cdot]$ represents the imaginary part of the variable. From $\beta_{im}\neq 0$ and (21), we have

$$\beta_{im}=\gamma_{im}(\tilde{\boldsymbol{h}}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\boldsymbol{b}_{im})^{*}=\gamma_{im}\boldsymbol{b}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\tilde{\boldsymbol{h}}_{im} \tag{22}$$

or

$$\tilde{\boldsymbol{h}}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\boldsymbol{b}_{im}=0, \tag{23}$$

where $\gamma_{im}\in\mathbb{R}\backslash\{0\}$. When (22) holds, we can derive a quadratic equation by substituting (22) into (20) as follows:

$$\boldsymbol{b}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\boldsymbol{b}_{im}|\boldsymbol{b}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\tilde{\boldsymbol{h}}_{im}|^2\gamma_{im}^2-|\tilde{\boldsymbol{h}}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\boldsymbol{b}_{im}|^2\gamma_{im}-1=0. \tag{24}$$

By substituting the solution $\gamma_{im}$ of (24) into (22), we have

$$\beta_{im}=-\frac{\boldsymbol{b}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\tilde{\boldsymbol{h}}_{im}}{2\boldsymbol{b}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\boldsymbol{b}_{im}}\left(-1\pm\sqrt{1+\frac{\boldsymbol{b}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\boldsymbol{b}_{im}}{|\tilde{\boldsymbol{h}}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\boldsymbol{b}_{im}|^2}}\right), \tag{25}$$

where the $\pm$ sign in (25) should be positive to make $\mathcal{L}(\boldsymbol{W})$ smaller. On the other hand, when (23) holds, the solution of (20) becomes

$$\beta_{im}=\frac{e^{\mathrm{j}\phi_{im}}}{\sqrt{\boldsymbol{b}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}^{-1}\boldsymbol{b}_{im}}}, \tag{26}$$

where $\phi_{im}\in(-\pi,\pi]$ denotes an arbitrary phase and j is the imaginary unit. Since $\phi_{im}$ does not change the value of $\mathcal{L}(\boldsymbol{W})$, $\phi_{im}$ is set to satisfy $e^{\mathrm{j}\phi_{im}}=(\det\boldsymbol{W}_i)/|\det\boldsymbol{W}_i|$. From (18), (25), (26), and the relation $\boldsymbol{b}_{im}=(\det\boldsymbol{W}_i)^*(\boldsymbol{W}_i^{-1})^{\mathrm{H}}\boldsymbol{e}_m$, we obtain the following update rules of $\tilde{\boldsymbol{w}}_{im}$:

$$\boldsymbol{u}_{im}\leftarrow(\boldsymbol{W}_i^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im})^{-1}\boldsymbol{e}_m, \tag{27}$$

$$\hat{\boldsymbol{u}}_{im}\leftarrow\tilde{\boldsymbol{Q}}_{im}^{-1}\tilde{\boldsymbol{h}}_{im}, \tag{28}$$

$$a_{im}\leftarrow\boldsymbol{u}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}\boldsymbol{u}_{im}, \tag{29}$$

$$\hat{a}_{im}\leftarrow\boldsymbol{u}_{im}^{\mathrm{H}}\tilde{\boldsymbol{Q}}_{im}\hat{\boldsymbol{u}}_{im}, \tag{30}$$

$$\tilde{\boldsymbol{w}}_{im}\leftarrow\begin{cases}\frac{\boldsymbol{u}_{im}}{\sqrt{a_{im}}}-\hat{\boldsymbol{u}}_{im} & (\hat{a}_{im}=0),\\ \frac{\hat{a}_{im}}{2a_{im}}\left[1-\sqrt{1+\frac{4a_{im}}{|\hat{a}_{im}|^2}}\right]\boldsymbol{u}_{im}-\hat{\boldsymbol{u}}_{im} & (\hat{a}_{im}\neq 0).\end{cases} \tag{31}$$

This update of $\tilde{\boldsymbol{w}}_{im}$ guarantees the monotonic nonincrease in $\mathcal{L}(\boldsymbol{W})$.

## IV. Experimental Evaluation

### A. Experimental Conditions

We confirmed the validity of the proposed columnwise update by conducting a music source separation task. We compared five methods: independent low-rank matrix analysis (ILRMA) [9], DNN with Wiener filtering (DNN+WF) [24], combination of full-rank spatial covariance model and DNN source model (FSCM+DNN) [15], conventional IDLMA with row-wise IP (Row-IDLMA), and the proposed IDLMA with the columnwise update (Column-IDLMA). Note that ILRMA is a "blind" (unsupervised) technique, but we show its performance just for reference to understand to what extent the supervised methods (DNN+WF, FSCM+DNN, Row-IDLMA, and Column-IDLMA) can improve the performance. For all methods except DNN+WF, we updated the spatial model 500 times. For FSCM+DNN, Row-IDLMA, and Column-IDLMA, the scale parameter matrix $\boldsymbol{R}_n$ was updated by $\mathrm{DNN}_n$ after every 10 iterations of the spatial parameter optimization.

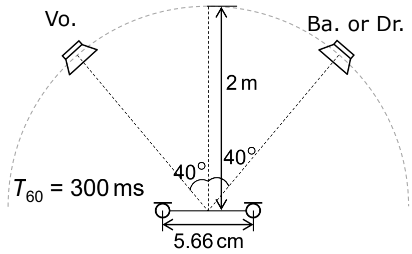


Fig. 3. Recording condition of impulse responses obtained from RWCP database.

We used the DSD100 dataset of SiSEC2016 [25] as the dry sources and the training dataset of DNN. The 50 songs in the `dev` data were used to train $\mathrm{DNN}_n$ and the top 25 songs in alphabetical order in the `test` data were used for performance evaluation. The test songs were trimmed only in the interval of 30 to 60 s. To simulate reverberant mixtures, we produced two-channel observed signals by convoluting the impulse response E2A ($T_{60}=300\,\mathrm{ms}$) obtained from the RWCP database [26] with each source, and mixtures of bass (Ba.) and vocal (Vo.) or drums (Dr.) and Vo. were created. The recording condition of E2A is shown in Fig. 3. All the signals were downsampled to $8\,\mathrm{kHz}$. An STFT was performed using a 512-ms-long Hamming window with a 256-ms-long shift. We used the signal-to-distortion ratio (SDR) [27] to evaluate the total separation performance.

In this paper, the number of hidden layers in the constructed fully connected DNN was set to four. Each layer had 1024 units, and a rectified linear unit was used for the output of each layer. To optimize the DNN, we added the term $(\lambda/2)\sum_q {g_q}^2$ to (11) for regularization, where $g_q$ is the weight coefficient in DNN, and ADADELTA [28] with a 128-size minibatch was performed for 2000 epochs. The parameter $\epsilon$ was experimentally optimized and set to $(0.1\times(IJ)^{-1}\sum_{i,j}{\sigma_{ijn}}^2)^{\frac{1}{2}}$. The other parameters were set to $\delta=10^{-5}$ and $\lambda=10^{-5}$.

### B. Results

Figs. 4 and 5 show the average SDR improvements for Ba./Vo. and Dr./Vo. separation, respectively. The proposed Column-IDLMA achieves the best SDR improvement among all the state-of-the-art blind and supervised methods in both Ba./Vo. and Dr./Vo separation. In particular, Column-IDLMA outperforms Row-IDLMA by over 0.8 dB in Ba./Vo. These results confirm that the proposed columnwise update algorithm is more appropriate than IP in IDLMA, where the accuracy of the scale parameter estimation by DNN will depend on the type of source.

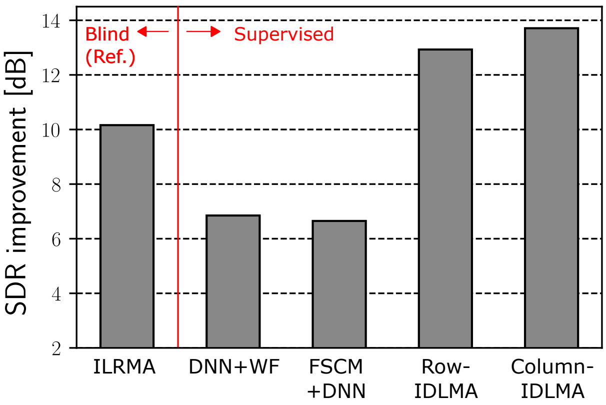


Fig. 4. Average SDR improvement of 25 Ba./Vo. songs.

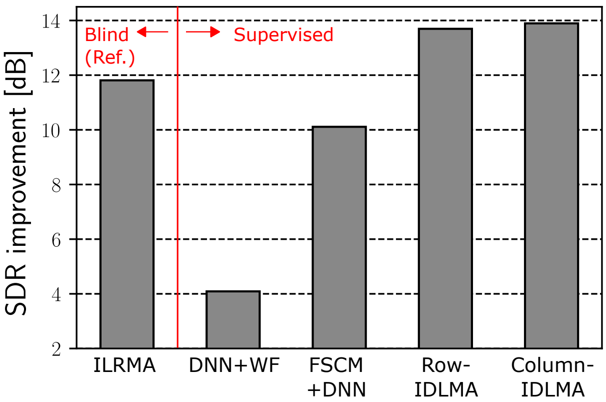


Fig. 5. Average SDR improvement of 25 Dr./Vo. songs.

## V. Conclusions

In this paper, we derived a new columnwise update algorithm of the demixing matrix in IDLMA, which simultaneously utilizes all source models for each update. Owing to this property, the proposed update algorithm is robust against the variance of the accuracy in the DNN inference w.r.t. the type of source. For this purpose, we employed VCD, which is a convergence-guaranteed algorithm applicable to the sum of the quadric, linear, and negative log-determinant terms. Experimental results showed that the proposed columnwise update algorithm has superior separation performance to the conventional row-wise update algorithm.

## Acknowledgment

This work was partly supported by SECOM Science and Technology Foundation and JSPS KAKENHI Grant Numbers JP17H06101, JP19H01116, and JP19K20306.

## References

[1] H. Sawada, N. Ono, H. kameoka, D. Kitamura, H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Transactions on Signal and Information Processing*, vol. 8, no. e12, pp. 1–14, 2019.

[2] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. SAP*, vol. 11, no. 2, pp. 109–116, 2003.

[3] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, pp. 666–678, 2006.

[4] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 70–79, 2007.

[5] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[6] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

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

[8] H. Kameoka, T. Yoshioka, M. Hamamura, J. L. Roux, and K. Kashino, "Statistical model of speech signals based on composite autoregressive system with application to blind source separation," in *Proc. LVA/ICA*, 2010, pp. 245–253.

[9] 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 Trans. ASLP*, vol. 14, no. 9, pp. 1626–1641, 2016.

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," in *Audio Source Separation*, S. Makino, Ed. Springer, Cham, 2018, ch. 6, pp. 125–155.

[11] D. Kitamura, S. Mogami, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Y. Takahashi, and K. Kondo, "Generalized independent low-rank matrix analysis using heavy-tailed distributions for blind source separation," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 28, pp. 1–25, 2018.

[12] R. Ikeshita and Y. Kawaguchi, "Independent low-rank matrix analysis based on multivariate complex exponential power distribution," in *Proc. ICASSP*, 2018, pp. 741–745.

[13] S. Mogami, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, K. Kondo, H. Nakajima, and N. Ono, "Independent low-rank matrix analysis based on time-variant sub-Gaussian source model," in *Proc. APSIPA*, 2018, pp. 1684–1691.

[14] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. ICASSP*, 2015, pp. 116–120.

[15] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.

[16] Y.-H. Tu, J. Du, L. Sun, and C.-H. Lee, "LSTM-based iterative mask estimation and post-processing for multi-channel speech enhancement," in *Proc. APSIPA*, 2017.

[17] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, "Deep learning based speech beamforming," in *Proc. ICASSP*, 2018, pp. 5389–5393.

[18] N. Makishima, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, K. Kondo, and H. Nakajima, "Generalized-Gaussian-distribution-based independent deeply learned matrix analysis for multichannel audio source separation," in *Proc. INTERNOISE*, no. 1260, 2019.

[19] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. ASLP*, 2019. (DOI: 10.1109/TASLP.2019.2925450)

[20] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2010.

[21] N. Makishima, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, and K. Kondo, "Column-wise update algorithm for independent deeply learned matrix analysis," in *Proc. ICA*, 2019, in press.

[22] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis

for multichannel audio source separation," in *Proc. EUSIPCO*, 2018, pp. 1571–1575.
[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 *Proc. ICASSP*, 2018, pp. 746–750.
[24] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. ICASSP*, 2015, pp. 2135–2139.
[25] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. LVA/ICA*, 2012, pp. 323–332.
[26] S. Nakamura, K. Hiyane, F. Asano, T. Nishimura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. LREC*, 2000, pp. 965–968.
[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.
[28] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *CoRR*, vol. abs/1212.5701, 2012.