EFFICIENT MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION EXPLOITING RANK-1 SPATIAL MODEL

Daichi Kitamura[1], *Nobutaka Ono*[2, 1], *Hiroshi Sawada*[3], *Hirokazu Kameoka*[3, 4], *Hiroshi Saruwatari*[4]

[1] The Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan
[2] National Institute of Informatics, Tokyo, Japan
[3] Nippon Telegraph and Telephone Corporation, Tokyo, Japan
[4] The University of Tokyo, Tokyo, Japan

ABSTRACT

This paper proposes a new efficient multichannel nonnegative matrix factorization (NMF) method. Recently, multichannel NMF (MNMF) has been proposed as a means of solving the blind source separation problem. This method estimates a mixing system of sources and attempts to separate them in a blind fashion. However, this method is strongly dependent on its initial values because there are no constraints in the spatial models. To solve this problem, we introduce a rank-1 spatial model into MNMF. The proposed method estimates a demixing matrix while representing sources using NMF bases and can be optimized by the update rules of independent vector analysis and conventional single-channel NMF. Experimental results show the efficacy of the proposed method in terms of robustness and convergence speed.

Index Terms— blind source separation, nonnegative matrix factorization, independent vector analysis, rank-1 spatial model

1. INTRODUCTION

Blind source separation (BSS) is a technique for separating specific sources from a recorded sound without any information about the recording environment, mixing system, or source locations. In a determined or overdetermined situation (number of microphones ≥ number of sources), independent component analysis (ICA) [1] is the method most commonly used to solve the BSS problem, and many ICA-based techniques have been proposed [2]–[6]. On the other hand, for an underdetermined situation (number of microphones < number of sources) including monaural recording, nonnegative matrix factorization (NMF) [7] has received much attention [8, 9]. BSS is generally used to solve speech separation problems, but recently the use of BSS for music signals has also become an active research area [10].

As a means of solving the permutation problem [11]–[13] in time-frequency domain ICA, independent vector analysis (IVA) [14] has been proposed. Such ICA-based methods assume independence between the sources in order to estimate a demixing matrix. However, if the observed signal frequently contains co-occurring sources, such as music signals, these methods cannot separate them with high accuracy because the independence between the sources is weakened.

In NMF-based methods, the decomposed bases (spectral patterns) must be clustered so as to represent the specific sources. One effective way of achieving this is to utilize a sample sound of the target signal [15, 16]. However, such supervision cannot be utilized in BSS. To solve this problem, multichannel NMF (MNMF) has been proposed [17]–[22]. In particular, MNMF methods [20]–[22] estimate a mixing system of the sources via spatial covariance matrices, which is utilized for source clustering.

In [17]–[19], an instantaneous mixture in the time domain is assumed, and only the observed gain of each microphone is modeled by extending NMF to nonnegative tensor factorization. For a convolutive mixture, we must treat complex values in the time-frequency domain. MNMF methods [20, 21] separately model the mixing system (as spatial covariance matrices) and the sources (using the NMF representation). These methods can be considered as a multichannel extension of NMF. Update rules based on the expectation-maximization algorithm have been derived. In [22], a unified MNMF scheme was proposed. This method employs Hermitian positive definite covariance matrices to model both the spatial and source components, and multiplicative update rules have been derived. However, it was reported that the algorithms [20]–[22] were sensitive to the initial values in source separation tasks.

In this paper, we propose a new efficient MNMF method with a rank-1 spatial model. Instead of estimating the mixing system, the proposed method estimates the demixing matrix while representing the sources using NMF bases. In addition, the proposed method can be optimized using the fast update rules of IVA and single-channel NMF. This fact enabled us to reveal the relationship between IVA and MNMF by analysis. The efficacy of the proposed method is confirmed experimentally.

2. CONVENTIONAL METHODS

2.1. Formulation

Let the numbers of sources and microphones (channels) be N and M, respectively. The multichannel source and the observed and separated signals in each time-frequency slot are described as

$$\boldsymbol{s}_{ij} = (s_{ij,1} \cdots s_{ij,N})^{\mathrm{t}}, \tag{1}$$

$$\boldsymbol{x}_{ij} = (x_{ij,1} \cdots x_{ij,M})^{\mathrm{t}}, \tag{2}$$

$$\boldsymbol{y}_{ij} = (y_{ij,1} \cdots y_{ij,N})^{\mathrm{t}}, \tag{3}$$

where $i = 1, \ldots, I$; $j = 1, \ldots, J$; $n = 1, \ldots, N$; and $m = 1, \ldots, M$ are the integral indexes of the frequency bins, time frames, sources, and channels, respectively, $^{\mathrm{t}}$ denotes the vector transpose, and all the entries of these vectors are complex values. When the window size in a short-time Fourier transform (STFT) is sufficiently long compared with the impulse responses, we can approximately represent the observed signal as

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

where $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \cdots \boldsymbol{a}_{i,N})$ is an $M \times N$ mixing matrix and $\boldsymbol{a}_{i,n}$ is the steering vector for each source. In the determined situation ($M = N$), we can define the demixing matrix $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1} \cdots \boldsymbol{w}_{i,M})^{\mathrm{h}}$, and the separated signal is represented as

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

where $^{\mathrm{h}}$ denotes the Hermitian transpose.

2.2. IVA

IVA is one of the techniques used to solve the permutation problem [13] and can be applied only in the overdetermined situation ($M \geq N$). For simplicity, let M be equal to N. In this method, we define the source component as a vector that consists of all frequency bins, given as

$$\boldsymbol{y}_{j,m} = (y_{1j,m} \cdots y_{Ij,m})^{\mathrm{t}}. \tag{6}$$

IVA can be used to estimate the demixing matrix $\boldsymbol{W}_i$ by assuming both independence between the sources (vectors) and a higher-order correlation between the frequency bins in each source [14]. The cost function of IVA is defined as follows:

$$Q_{\mathrm{IVA}}(\boldsymbol{W}) = \textstyle\sum_m \frac{1}{J} \sum_j G(\boldsymbol{y}_{j,m}) - \sum_i \log |\det \boldsymbol{W}_i|, \tag{7}$$

where J is the number of time frames and $G(\boldsymbol{y}_{j,m})$ is a contrast function. When $\boldsymbol{y}_{j,m}$ obeys a probabilistic density function $p(\boldsymbol{y}_{j,m})$, the contrast function $G(\boldsymbol{y}_{j,m})$ is given as $-\log p(\boldsymbol{y}_{j,m})$. In IVA, $G(\boldsymbol{y}_{j,m}) = \|\boldsymbol{y}_{j,m}\|_2$ is often used [14], which assumes a spherical Laplace distribution for the source prior, where $\|\cdot\|_2$ denotes the L_2 norm. For the minimization of (7), fast and stable update rules, which are derived by an auxiliary function technique, have been proposed [23].

2.3. MNMF

MNMF is a natural extension of simple NMF for multichannel signals [22]. This method can be applied even in the underdetermined situation ($M < N$). The observed signal is represented as

$$\mathsf{X}_{ij} = \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{h}}, \tag{8}$$

where X_{ij} is a Hermitian positive definite matrix of size $M \times M$. The diagonal elements of X_{ij} represent real-valued powers observed by the microphone, and the nondiagonal elements represent the complex-valued correlations between the microphones. The decomposition model of MNMF is expressed as

$$\mathsf{X}_{ij} \approx \hat{\mathsf{X}}_{ij} = \textstyle\sum_k \left(\sum_n \mathsf{H}_{i,n} z_{nk}\right) t_{ik} v_{kj}, \tag{9}$$

where $k = 1, \ldots, K$ is the integral index of the spectral bases, and $\mathsf{H}_{i,n}$ is an $M \times M$ Hermitian positive definite matrix, which comprises the spatial covariance for each frequency i and source n. In addition, $z_{nk} \in \mathbb{R}_{[0,1]}$ is a latent variable that indicates whether the kth basis belongs to the nth source ($z_{nk} = 1$) or not ($z_{nk} = 0$) and satisfies $\sum_n z_{nk} = 1$; $t_{ik} \in \mathbb{R}_{\geq 0}$ and $v_{kj} \in \mathbb{R}_{\geq 0}$ are the elements of the basis matrix $\boldsymbol{T} \in \mathbb{R}_{\geq 0}^{I \times K}$ and activation matrix $\boldsymbol{V} \in \mathbb{R}_{\geq 0}^{K \times J}$. Figure 1 shows a conceptual model of MNMF. In BSS, the mixing and demixing systems are unknown. MNMF decomposes the observed signal into $\boldsymbol{T}$ and $\boldsymbol{V}$ and simultaneously optimizes the spatial covariance matrices H that correspond to each sources. Then, the sources are separated by associating these variables $\boldsymbol{T}$ and $\boldsymbol{V}$ with H by using a cluster-indicator latent variable $\boldsymbol{Z} \in \mathbb{R}_{\geq 0}^{N \times K}$. The cost function based on Itakura-Saito divergence is defined as [22]

$$Q_{\mathrm{MNMF}} = \textstyle\sum_{i,j} \left[\mathrm{tr}(\mathsf{X}_{ij} \hat{\mathsf{X}}_{ij}^{-1}) + \log \det \hat{\mathsf{X}}_{ij} \right], \tag{10}$$

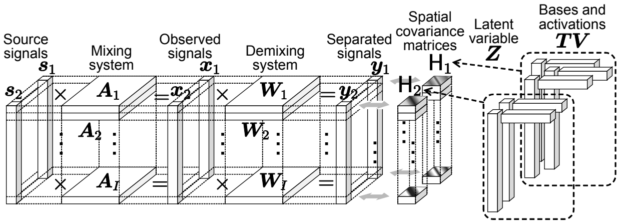

Fig. 1. Conceptual model of MNMF ($N = M = 2$).

where constant terms are omitted. Update rules to minimize (10) have been derived by the auxiliary function technique [22].

3. PROPOSED METHOD

3.1. Motivation and strategy

The separation performance of IVA is degraded for music signals because the independence between the sources is weakened by the frequent co-occurrence. In addition, IVA assumes that all the frequency bins have the same time-varying gains. Therefore, IVA is not suitable for signals that have harmonic structures. MNMF has problems in terms of the convergence speed and its strong dependence on its initial values. This is because a huge number of variables should be optimized using only one cost function, and because there is no constraint for optimizing the spatial covariance matrices H.

To solve these problems, we propose a new efficient MNMF method that employs a rank-1 approximation for the spatial covariance matrices H. In this approach, similarly to in IVA, we assume the determined situation and the linear time-invariant mixing system described by (4) to restrict the flexibility of the spatial model. In addition, instead of estimating H, the proposed method estimates the demixing matrix $\boldsymbol{W}_i$ while representing the sources using NMF bases.

3.2. Derivation of cost function

If we assume that the mixing system shown in Fig. 1 is represented by the mixing matrix $\boldsymbol{A}_i = (\boldsymbol{a}_{i,1} \cdots \boldsymbol{a}_{i,n})$ appearing in (4), the spatial covariance matrix $\mathsf{H}_{i,n}$ can be approximated by a rank-1 matrix that is an outer product of the steering vector $\boldsymbol{a}_{i,n}$ as follows:

$$\mathsf{H}_{i,n} = \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{h}}. \tag{11}$$

To introduce the rank-1 approximation into MNMF, we substitute (11) into (9) and reformulate $\hat{\mathsf{X}}_{ij}$ using the mixing matrix $\boldsymbol{A}_i$ as follows:

$$\begin{aligned}\hat{\mathsf{X}}_{ij} &= \textstyle\sum_k \left(\sum_n \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{h}} z_{nk}\right) t_{ik} v_{kj} \\ &= \textstyle\sum_n \boldsymbol{a}_{i,n} \boldsymbol{a}_{i,n}^{\mathrm{h}} \sum_k z_{nk} t_{ik} v_{kj} \\ &= \boldsymbol{A}_i \mathsf{D}_{ij} \boldsymbol{A}_i^{\mathrm{h}},\end{aligned} \tag{12}$$

where

$$\mathsf{D}_{ij} = \mathrm{diag}\left(d_{ij,1}, \ldots, d_{ij,N}\right), \tag{13}$$

$$d_{ij,n} = \textstyle\sum_k z_{nk} t_{ik} v_{kj}. \tag{14}$$

By substituting (12) into the MNMF cost function (10), we obtain

$$Q = \textstyle\sum_{i,j} \left[\mathrm{tr}\left(\boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{h}} \left(\boldsymbol{A}_i^{\mathrm{h}}\right)^{-1} \mathsf{D}_{ij}^{-1} \boldsymbol{A}_i^{-1} \right) + \log \det \boldsymbol{A}_i \mathsf{D}_{ij} \boldsymbol{A}_i^{\mathrm{h}} \right]. \tag{15}$$

Table 1. Music sources

ID	Song	Source (1/2/3)
1	bearlin-roads__snip_85_99	acoustic_guit_main/bass/vocals
2	fort_minor-remember_the_name__snip_54_78	drums/violins_synth/vocals
3	ultimate_nz_tour__snip_43_61	guitar/synth/vocals

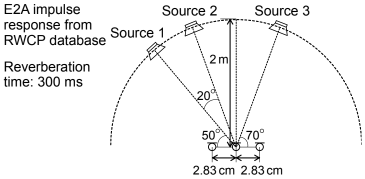
E2A impulse response from RWCP database
Reverberation time: 300 ms
Source 1
Source 2
Source 3
2 m
20°
50°
70°
2.83 cm
2.83 cm

Fig. 2. Recording conditions of room impulse response.

In the overdetermined situation (we let M equal N for simplicity), the demixing matrix $\boldsymbol{W}_i$ exists and we can transform the variables, i.e., the observed signal $\boldsymbol{x}_{ij}$ and the mixing matrix $\boldsymbol{A}_i$, to the separated signal $\boldsymbol{y}_{ij} = \boldsymbol{W}_i \boldsymbol{x}_{ij}$ and the demixing matrix $\boldsymbol{W}_i = \boldsymbol{A}_i^{-1}$ respectively, as follows:

$$\begin{aligned}
Q &= \textstyle\sum_{i,j} \Big[\mathrm{tr}\Big(\boldsymbol{W}_i^{-1}\boldsymbol{y}_{ij}\boldsymbol{y}_{ij}^{\mathrm{h}}\left(\boldsymbol{W}_i^{\mathrm{h}}\right)^{-1}\boldsymbol{W}_i^{\mathrm{h}}\mathsf{D}_{ij}^{-1}\boldsymbol{W}_i\Big) \\
&\qquad\qquad + \log\left(\det \boldsymbol{A}_i\right)\left(\det \mathsf{D}_{ij}\right)\left(\det \boldsymbol{A}_i^{\mathrm{h}}\right)\Big] \\
&= \textstyle\sum_{i,j} \Big[\mathrm{tr}\Big(\boldsymbol{W}_i\boldsymbol{W}_i^{-1}\boldsymbol{y}_{ij}\boldsymbol{y}_{ij}^{\mathrm{h}}\left(\boldsymbol{W}_i^{\mathrm{h}}\right)^{-1}\boldsymbol{W}_i^{\mathrm{h}}\mathsf{D}_{ij}^{-1}\Big) \\
&\qquad\qquad + 2\log|\det \boldsymbol{A}_i| + \log\det \mathsf{D}_{ij}\Big] \\
&= \textstyle\sum_{i,j} \Big[\mathrm{tr}\left(\boldsymbol{y}_{ij}\boldsymbol{y}_{ij}^{\mathrm{h}}\mathsf{D}_{ij}^{-1}\right) - 2\log|\det \boldsymbol{W}_i| + \sum_m \log d_{ij,m}\Big] \\
&= \textstyle\sum_{i,j} \left[\sum_m \dfrac{|y_{ij,m}|^2}{\sum_k z_{mk}t_{ik}v_{kj}} - 2\log|\det \boldsymbol{W}_i| + \sum_m \log \sum_k z_{mk}t_{ik}v_{kj}\right].
\end{aligned} \tag{16}$$

In the conventional MNMF, the separated signal is obtained by clustering H, $\boldsymbol{T}$, and $\boldsymbol{V}$ into specific sources using a latent variable $\boldsymbol{Z}$. The proposed method estimates the demixing matrix $\boldsymbol{W}_i$ to obtain the separated signal $\boldsymbol{y}_{ij}$, where we approximately decompose $\boldsymbol{y}_{ij}$ into z_{mk}, t_{ik}, and v_{kj} in each iteration.

3.3. Relationship between IVA and MNMF

The first and second terms in the cost function (16) are equivalent to the IVA cost function (7), and the first and third terms in (16) are equivalent to a single-channel NMF cost function described as

$$Q_{\mathrm{NMF}} = \textstyle\sum_{i,j}\left[\dfrac{|y_{ij}|^2}{\sum_l t_{il}v_{lj}} + \log \sum_l t_{il}v_{lj}\right], \tag{17}$$

where $l = 1, \ldots, L$ indicates the basis index. This fact reveals the relationship between IVA and MNMF, namely, MNMF with a rank-1 approximation, which assumes a linear time-invariant mixing system in the time-frequency domain, is essentially equivalent to IVA with a basis decomposition model. Therefore, the proposed method can be considered as an intermediate model between IVA and MNMF in terms of the degree of freedom. From the IVA side, we introduced the basis decomposition model with NMF to capture the actual spectral patterns, and from the MNMF side, an approximation for the spatial model was introduced to make the optimization more efficient. However, the source priors of IVA and the proposed method are different. IVA generally assumes the spherical Laplace distribution, which has the same variance for all frequency bins, as the source prior by setting $G(\boldsymbol{y}_{j,m}) = \|\boldsymbol{y}_{j,m}\|_2$ in (7). The proposed method with (16) assumes independent complex Gaussian distributions in each time-frequency slot [24], similarly to conventional MNMF.

Table 2. Experimental conditions

Sampling frequency	Downsampled from 44.1 kHz to 16 kHz
FFT length	512 ms
Window shift	128 ms
Number of bases K	Proposed method 1: $L = 30$ ($K = 90$) MNMF+MWF and Proposed method 2: $K = 90$
Number of iterations	200

3.4. Update rules

If we eliminate z_{mk} in (16), the differentials $\partial Q/\partial t_{ik}$ and $\partial Q/\partial v_{kj}$ become identical to $\partial Q_{\mathrm{NMF}}/\partial t_{il}$ and $\partial Q_{\mathrm{NMF}}/\partial v_{lj}$, respectively. Therefore, when $z_{mk} \in \{0, 1\}$ and all the sound sources are modeled by the same number of bases L (namely, $L \times M = K$), the update rules of $t_{il,m}$ and $v_{lj,m}$ are the same as those of single-channel NMF, i.e.,

$$t_{il,m} \leftarrow t_{il,m}\sqrt{\frac{\sum_j |y_{ij,m}|^2 v_{lj,m}\left(\sum_{l'} t_{il',m}v_{l'j,m}\right)^{-2}}{\sum_j v_{lj,m}\left(\sum_{l'} t_{il',m}v_{l'j,m}\right)^{-1}}}, \tag{18}$$

$$v_{lj,m} \leftarrow v_{lj,m}\sqrt{\frac{\sum_i |y_{ij,m}|^2 t_{il,m}\left(\sum_{l'} t_{il',m}v_{l'j,m}\right)^{-2}}{\sum_i t_{il,m}\left(\sum_{l'} t_{il',m}v_{l'j,m}\right)^{-1}}}, \tag{19}$$

$$r_{ij,m} = \textstyle\sum_l t_{il,m}v_{lj,m}, \tag{20}$$

where $r_{ij,m}$ is an estimated variance of each source [24]. Alternatively, if we employ a continuous-valued z_{mk} to cluster the bases into specific sources, we can derive the following update rules of z_{mk}, t_{ik}, and v_{kj} by minimizing (16) by the auxiliary function technique;

$$z_{mk} \leftarrow z_{mk}\sqrt{\frac{\sum_{i,j} |y_{ij,m}|^2 t_{ik}v_{kj}\left(\sum_{k'} z_{mk'}t_{ik'}v_{k'j}\right)^{-2}}{\sum_{i,j} t_{ik}v_{kj}\left(\sum_{k'} z_{mk'}t_{ik'}v_{k'j}\right)^{-1}}}, \tag{21}$$

$$t_{ik} \leftarrow t_{ik}\sqrt{\frac{\sum_{j,m} |y_{ij,m}|^2 z_{mk}v_{kj}\left(\sum_{k'} z_{mk'}t_{ik'}v_{k'j}\right)^{-2}}{\sum_{j,m} z_{mk}v_{kj}\left(\sum_{k'} z_{mk'}t_{ik'}v_{k'j}\right)^{-1}}}, \tag{22}$$

$$v_{kj} \leftarrow v_{kj}\sqrt{\frac{\sum_{i,m} |y_{ij,m}|^2 z_{mk}t_{ik}\left(\sum_{k'} z_{mk'}t_{ik'}v_{k'j}\right)^{-2}}{\sum_{i,m} z_{mk}t_{ik}\left(\sum_{k'} z_{mk'}t_{ik'}v_{k'j}\right)^{-1}}}, \tag{23}$$

$$r_{ij,m} = \textstyle\sum_k z_{mk}t_{ik}v_{kj}, \tag{24}$$

where we calculate $z_{mk} \leftarrow z_{mk}/\sum_{m'} z_{m'k}$ to ensure $\sum_m z_{mk} = 1$ at each iteration.

The demixing matrix $\boldsymbol{W}_i$ can be optimized by the update rule of IVA because the differentials $\partial Q/\partial \boldsymbol{W}_i$ and $\partial Q_{\mathrm{IVA}}/\partial \boldsymbol{W}_i$ are equivalent. Fast and stable update rules for IVA have been derived as [23]

$$V_{i,m} = \frac{1}{J}\textstyle\sum_j \dfrac{1}{r_{ij,m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^{\mathrm{h}}, \tag{25}$$

$$\boldsymbol{w}_{i,m} \leftarrow \left(\boldsymbol{W}_i V_{i,m}\right)^{-1}\boldsymbol{e}_m, \tag{26}$$

$$\boldsymbol{w}_{i,m} \leftarrow \boldsymbol{w}_{i,m}\left(\boldsymbol{w}_{i,m}^{\mathrm{h}}V_{i,m}\boldsymbol{w}_{i,m}\right)^{-\frac{1}{2}}, \tag{27}$$

$$y_{ij,m} \leftarrow \boldsymbol{w}_{i,m}^{\mathrm{h}}\boldsymbol{x}_{ij}, \tag{28}$$

where $\boldsymbol{e}_m$ denotes the unit vector with the mth element equal to unity.

We estimate all the variables that minimize (16) by iterating (18)–(20) or (21)–(24) and (25)–(28) alternately. Note that the scale

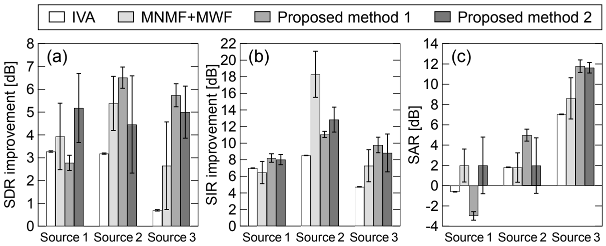

Fig. 3. Average scores for ID1 data: (a) SDR improvement, (b) SIR improvement, and (c) SAR.

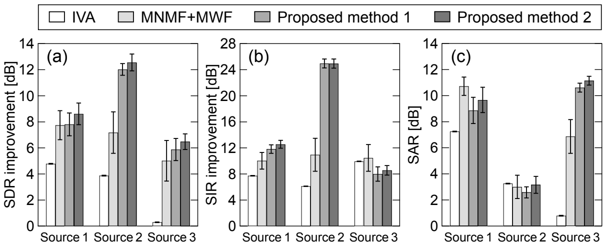

Fig. 4. Average scores for ID2 data: (a) SDR improvement, (b) SIR improvement, and (c) SAR.

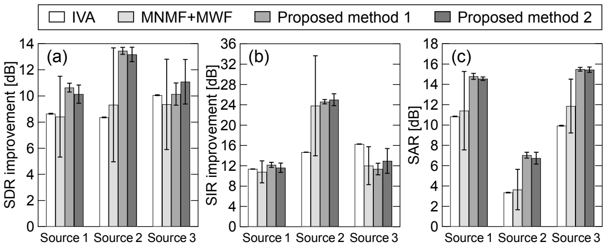

Fig. 5. Average scores for ID3 data: (a) SDR improvement, (b) SIR improvement, and (c) SAR.

ambiguity exists between $\boldsymbol{W}_i$ and $r_{ij,m}$ because both of them can determine the scale. Therefore, the estimated variance $r_{ij,m}$ has a risk of diverging. To avoid this problem, we normalize $\boldsymbol{W}_i$ and $r_{ij,m}$ at each iteration. The signal scale can be restored by applying a back-projection technique [12] after the optimization.

4. EXPERIMENT

4.1. Conditions

To confirm the efficacy of the proposed method, we conducted an evaluation experiment using professional music signals. In this experiment, we compared four methods, namely, IVA [23], MNMF with a multichannel Wiener filter (referred to as MNMF+MWF) [22], Proposed method 1 (update using (18)–(20) and (25)–(28)), and Proposed method 2 (update using (21)–(24) and (25)–(28)). Proposed method 1 models all the sources with the same fixed number of bases, L. In Proposed method 2, we only set the total number of bases, K, and the sources are flexibly modeled with the optimal number of bases using the cluster-indicator $\boldsymbol{Z}$. We used three musics obtained from SiSEC [25] and selected three sources, as shown in Table 1. In addition, the observed signal, which has three channels, was created by convoluting the impulse response E2A (see Fig. 2) from RWCP database [26] with each source. The other experimental conditions are described in Table 2. As the evaluation scores, we used the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and sources-to-artifacts ratio (SAR) defined in [27]. SDR indicates the total separation performance, SIR indicates the degree of separation, and SAR indicates the absence of artificial distortion.

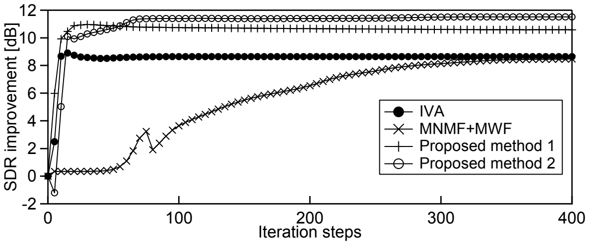

Fig. 6. Example of SDR convergence: result for Source 1 in ID3.

Table 3. Computational times for separation of ID1 (s)

IVA	MNMF+MWF	Proposed method 1	Proposed method 2
91.6	4498.4	121.0	173.4

4.2. Results

Figures 3–5 show the average scores and their deviations in 10 trials with various initializations. From these results, we confirm that IVA cannot achieve satisfactory separation because this method is not suitable for music signals. MNMF+MWF gives slightly better performance than IVA, and its maximum scores are comparable to the average scores of the proposed methods. However, owing to the lack of robustness, the error bars are relatively large. The proposed methods achieve good and stable performance, which is particularly evident in Figs. 4 and 5. This is because (a) the spectral patterns are effectively modeled by the decomposition of bases compared with IVA, and (b) the optimization of the demixing matrix using the IVA update rules results in a stable separation performance.

Figure 6 shows an example of SDR convergence for each method. Both IVA and the proposed methods show much faster convergence and better results than MNMF+MWF. The actual computational times in this experiment are shown in Table 3, where the calculations were performed using MATLAB 8.3 (64-bit) with an Intel Core i7-4790 (3.60 GHz) CPU. The computational times of the proposed methods are less than twice that of IVA. MNMF requires a longer computational time because of the eigenvalue decomposition for each $\mathsf{H}_{i,n}$.

5. CONCLUSION

This paper presents an efficient MNMF method that includes a rank-1 approximation of the covariance matrix. The proposed method can be optimized using the fast update rules of IVA and single-channel NMF. Also, we revealed that MNMF with the rank-1 approximation is essentially equivalent to IVA with the basis decomposition model. The experimental results show that the proposed method achieves faster convergence and better results than the conventional methods.

6. ACKNOWLEDGMENT

This work was partially supported by Grant-in-Aid for JSPS Fellows Grant Number 26·10796.

7. REFERENCES

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

[2] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol.22, pp.21–34, 1998.

[3] 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. Speech and Audio Processing*, vol.11, no.2, pp.109–116, 2003.

[4] H. Sawada, R. Mukai, S. Araki and S. Makino, "Convolutive blind source separation for more than two sources in the frequency domain," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.Ⅲ-885–Ⅲ-888, 2004.

[5] H. Buchner, R. Aichner and W. Kellerman, "A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics," *IEEE Trans. Speech and Audio Processing*, vol.13, no.1, pp.120–134, 2005.

[6] 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. Audio, Speech, and Language Processing*, vol.14, no.2, pp.666–678, 2006.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Proc. Advances in Neural Information Processing Systems*, vol.13, pp.556–562, 2001.

[8] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Processing*, vol.15, no.3, pp.1066–1074, 2007.

[9] A. Ozerov, C. Févotte and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.121–124, 2009.

[10] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino and S. Sagayama, "Constrained and regularized variants of nonnegative matrix factorization incorporating music-specific constraints," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.5365–5368, 2012.

[11] S. Kurita, H. Saruwatari, S. Kajita K. Takeda and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.5, pp.3140–3143, 2000.

[12] N. Murata, S. Ikeda and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol.41, no.1–4, pp.1–24, 2001.

[13] H. Sawada, R. Mukai, S. Araki and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Processing*, vol.12, no.5, pp.530–538, 2004.

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

[15] P. Smaragdis, B. Raj and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. International Conference on Independent Component Analysis and Signal Separation*, pp.414–421, 2007.

[16] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi and K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol.E97-A, no.5, pp.1113–1118, 2014.

[17] D. FitzGerald, M. Cranitch and E. Coyle, "Non-negative tensor factorisation for sound source separation," *Proc. Irish Signals and Systems Conference*, pp.8–12, 2005.

[18] R. M. Parry and I. A. Essa, "Estimating the spatial position of spectral components in audio," *Proc. International Conference on Independent Component Analysis and Blind Source Separation*, pp.666–673, Springer, Berlin and Heidelberg, 2006.

[19] Y. Mitsufuji and A. Roebel, "Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.71–75, 2013.

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

[21] S. Arberet, A. Ozerov, N.Q.K. Duong, E. Vincent, R. Gribonval, R. Bimbot and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," *Proc. Information Sciences Signal Processing and their Applications*, pp.1–4, 2010.

[22] 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 Processing*, vol.21, no.5, pp.971–982, 2013.

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

[24] C. Févotte, N. Bertin and J.-L. Durrieu "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol.21, no.3, pp.793–830, 2009.

[25] S. Araki, F. Nesta, E. Vincent, Z. Koldovský, G. Nolte, A. Ziehe and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation," *Proc. Latent Variable Analysis and Signal Separation*, pp.414–422, 2012.

[26] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. International Conference on Language Resources and Evaluation*, pp.965–968, 2000.

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