PAPER

Supervised Audio Source Separation Based on Nonnegative Matrix Factorization with Cosine Similarity Penalty

Yuta IWASE[†], *Nonmember and* **Daichi KITAMURA**[†a)], *Member*

SUMMARY In this study, we aim to improve the performance of audio source separation for monaural mixture signals. For monaural audio source separation, semisupervised nonnegative matrix factorization (SNMF) can achieve higher separation performance by employing small supervised signals. In particular, penalized SNMF (PSNMF) with orthogonality penalty is an effective method. PSNMF forces two basis matrices for target and nontarget sources to be orthogonal to each other and improves the separation accuracy. However, the conventional orthogonality penalty is based on an inner product and does not affect the estimation of the basis matrix properly because of the scale indeterminacy between the basis and activation matrices in NMF. To cope with this problem, a new PSNMF with cosine similarity between the basis matrices is proposed. The experimental comparison shows the efficacy of the proposed cosine similarity penalty in supervised audio source separation.

key words: *audio source separation, nonnegative matrix factorization, orthogonality, cosine similarity*

1. Introduction

Audio source separation is a technique of separating or extracting individual audio sources from the observed mixture signal. In particular, nonnegative matrix factorization (NMF) [1], [2] has been utilized in many situations [3]–[7], depending on the recording conditions and applications. As shown in Fig. 1, NMF is an algorithm that decomposes an observed nonnegative matrix $\boldsymbol{X}$ into two nonnegative matrices, a basis matrix $\boldsymbol{F}$ and an activation matrix $\boldsymbol{Q}$. The basis matrix includes frequently appearing spectral patterns in the observed matrix as basis vectors, and their time-varying gains are included in the activation matrix. Audio source separation can be achieved by clustering these estimated components into each source.

In recent years, deep learning methods have attracted wide attention and achieved high performance even in the source separation field. However, when the training data of target sources are excessively limited, the performance of deep-learning-based approaches degrades markedly. In such a situation, supervised or semisupervised NMF [8], [9] is still a promising algorithm, in which spectral patterns of the target sources are pretrained using only one sequence of their sample sounds, e.g., octave notes of the target instruments. In this paper, we only focus on a problem of music source separation based on SNMF, where only few sample notes of the target instruments are available as a supervision.

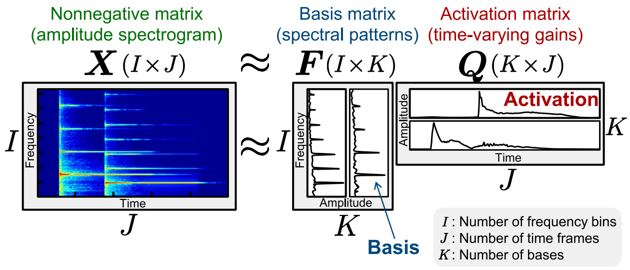
Nonnegative matrix (amplitude spectrogram)
Basis matrix (spectral patterns)
Activation matrix (time-varying gains)
$\boldsymbol{X}\,(I\times J)$
≈
$\boldsymbol{F}\,(I\times K)$
$\boldsymbol{Q}\,(K\times J)$
I
Frequency
Time
J
≈ I
Frequency
Amplitude
K
Amplitude
Activation
Time
J
K
Basis
I : Number of frequency bins
J : Number of time frames
K : Number of bases

Fig. 1 Matrix decomposition by NMF. Amplitude spectrogram of audio signal is input as nonnegative matrix.

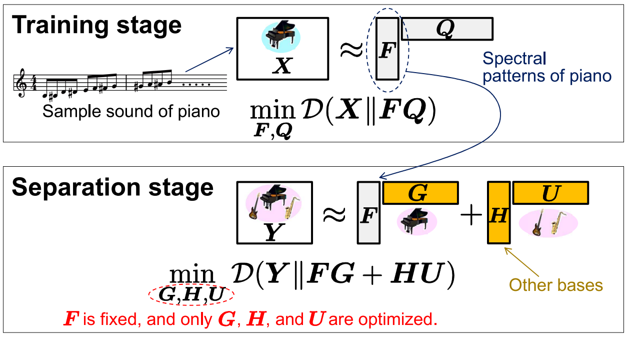
Training stage
X
≈
F
Q
Sample sound of piano
Spectral patterns of piano
$\min_{\boldsymbol{F},\boldsymbol{Q}} \mathcal{D}(\boldsymbol{X}\|\boldsymbol{F}\boldsymbol{Q})$
Separation stage
Y
≈
F
G
+
H
U
$\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{D}(\boldsymbol{Y}\|\boldsymbol{F}\boldsymbol{G}+\boldsymbol{H}\boldsymbol{U})$
Other bases
F is fixed, and only G, H, and U are optimized.

Fig. 2 SNMF algorithm. In training stage, sample sound of target source is decomposed by simple NMF, and supervised basis matrix $\boldsymbol{F}$ is obtained, which represents spectral patterns of target source. In separation stage, mixture sound is decomposed while $\boldsymbol{F}$ is fixed.

Figure 2 shows the SNMF-based source separation algorithm that consists of training and separation stages. In the training stage, an amplitude spectrogram of the sample sound of the target source, $\boldsymbol{X}$, is decomposed by simple NMF, and the supervised basis matrix $\boldsymbol{F}$ is obtained. In the separation stage, an amplitude spectrogram of the observed mixture $\boldsymbol{Y}$ is decomposed into the target spectrogram $\boldsymbol{F}\boldsymbol{G}$ and the other (nontarget) spectrogram $\boldsymbol{H}\boldsymbol{U}$ using the fixed supervised basis matrix $\boldsymbol{F}$.

When the target and nontarget sources in the observed mixture contain similar spectra, the separation performance of SNMF degrades. This is due to the fact that similar spectra can be represented by either the supervised basis matrix $\boldsymbol{F}$ or the nontarget basis matrix $\boldsymbol{H}$. In this case, either of the following problems occur: (a) some components of the target source are incorrectly included in $\boldsymbol{H}\boldsymbol{U}$, or (b) some components of the nontarget sources are inappropriately captured by $\boldsymbol{F}\boldsymbol{G}$. Since SNMF utilizes the sample sound of only the

Manuscript received November 12, 2021.
Manuscript publicized December 8, 2021.

[†] The authors are with the National Institute of Technology, Kagawa College, Takamatsu-shi, 761-8058 Japan.
a) E-mail: kitamura-d@t.kagawa-nct.ac.jp

DOI: 10.1587/transfun.2021EAP1149

target source, there exists a trade-off between the problems (a) and (b).

Penalized SNMF (PSNMF) [9] is a technique that addresses the above-mentioned trade-off problem. PSNMF forces the nontarget basis matrix $\boldsymbol{H}$ to be orthogonal to the supervised basis matrix $\boldsymbol{F}$ by adding an orthogonality penalty in the optimization of the separation stage. This is direct solution for the problem (a). We can avoid lack of the target source components in the estimated matrix $\boldsymbol{FG}$, although the problem (b) might be encouraged. Since semisupervised techniques aim to accurately extract the target source from a mixture, ensuring sound quality of the estimated target source is an important objective for many applications, e.g., music editing software. In addition, by tuning intensity of the orthogonality penalty, PSNMF can control the trade-off between the problems (a) and (b), which enables us to apply audio source separation to various objectives.

The penalty term proposed in [9] is defined as $\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{Fr}}^2$, a sum of the inner products between all the bases in $\boldsymbol{F}$ and $\boldsymbol{H}$ ($\|\cdot\|_{\mathrm{Fr}}$ is the Frobenius norm). Since the spectral patterns in $\boldsymbol{H}$ become dissimilar from those in $\boldsymbol{F}$, the separation performance is greatly improved. However, the formulation of conventional PSNMF is incorrect because the orthogonality penalty $\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{Fr}}^2$ does not affect the estimation of $\boldsymbol{H}$ properly owing to the scale indeterminacy between the basis and activation matrices in NMF.

In this paper, to solve the above-mentioned problem, two new orthogonality penalties based on cosine similarity are introduced to PSNMF. Since cosine similarity does not depend on scales of vectors, the proposed methods can appropriately force the nontarget spectral bases in $\boldsymbol{H}$ to be orthogonal to those in $\boldsymbol{F}$ even if the scale indeterminacy exists between $\boldsymbol{H}$ and $\boldsymbol{U}$. A convergence-guaranteed multiplicative update rule of the proposed methods is derived on the basis of the majorization-minimization (MM) algorithm [10]. The validity of the proposed methods is confirmed by music source separation experiments.

2. Conventional Methods

2.1 NMF and SNMF

The optimization problem in NMF [1], [2] is formulated as

$$\min_{\boldsymbol{F},\boldsymbol{Q}} \mathcal{D}(\boldsymbol{X}\|\boldsymbol{FQ}) \;\; \text{s.t.} \;\; f_{ik}, q_{kj'} \geq 0 \;\; \forall i,\; j',\; k, \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{I\times J'}$ is a nonnegative observed matrix and is an amplitude spectrogram in this paper. In addition, $\boldsymbol{F} \in \mathbb{R}_{\geq 0}^{I\times K}$ and $\boldsymbol{Q} \in \mathbb{R}_{\geq 0}^{K\times J'}$ are the basis and activation matrices, f_{ik} and $q_{kj'}$ are the elements of $\boldsymbol{F}$ and $\boldsymbol{Q}$, and $i = 1,2,\ldots,I$, $j' = 1,2,\ldots,J'$, and $k = 1,2,\ldots,K$ represent the indices of frequency bins, time frames, and basis vectors, respectively. Moreover, $\mathcal{D}(\boldsymbol{M}\|\boldsymbol{N})$ is a divergence function between two input matrices $\boldsymbol{M} \in \mathbb{R}_{\geq 0}^{I\times J}$ and $\boldsymbol{N} \in \mathbb{R}_{\geq 0}^{I\times J}$. In this paper, we only consider the generalized Kullback–Leibler (KL) divergence because it is experimentally confirmed that the KL-divergence-based NMF provides the highest performance for NMF-based audio source separation [9], [11]. The KL divergence is defined as

$$\mathcal{D}(\boldsymbol{M}\|\boldsymbol{N}) = \sum_{i,j}\left(m_{ij}\log\frac{m_{ij}}{n_{ij}} - m_{ij} + n_{ij}\right), \tag{2}$$

where m_{ij} and n_{ij} are the elements of matrices $\boldsymbol{M}$ and $\boldsymbol{N}$, respectively, and $j = 1,2,\ldots,J$. Thus, $\boldsymbol{F}$ and $\boldsymbol{Q}$ can be estimated by solving the minimization problem of Eq. (1).

In SNMF [8], the supervised basis matrix $\boldsymbol{F}$ is pretrained by applying simple NMF to the sample signal $\boldsymbol{X}$ of the target audio source in the training stage. The basis matrix $\boldsymbol{F}$ contains the frequently appearing spectral patterns of the target source as K column vectors (bases). In the separation stage, the amplitude spectrogram of the mixture signal $\boldsymbol{Y} \in \mathbb{R}_{\geq 0}^{I\times J}$ is decomposed using the supervised basis matrix $\boldsymbol{F}$ as follows:

$$\begin{aligned}&\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{D}(\boldsymbol{Y}\|\boldsymbol{FG}+\boldsymbol{HU})\\ &\text{s.t.}\;\; g_{kj}, h_{il}, u_{lj} \geq 0 \;\; \forall i,\; j,\; k,\; l,\end{aligned} \tag{3}$$

where $\boldsymbol{G} \in \mathbb{R}_{\geq 0}^{K\times J}$ is the activation matrix for $\boldsymbol{F}$. Moreover, $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{I\times L}$ and $\boldsymbol{U} \in \mathbb{R}_{\geq 0}^{L\times J}$ are the basis and activation matrices for representing nontarget sources, and g_{kj}, h_{il}, and u_{lj} are the elements of $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$, respectively, and $l = 1,2,\ldots,L$ represents the index of bases in $\boldsymbol{H}$. Ideally, the components of the target source in $\boldsymbol{Y}$ are extracted as $\boldsymbol{FG}$, and the components of the other nontarget sources are modeled by $\boldsymbol{HU}$, resulting in the separation of target and notarget sources. However, when the target and nontarget sources contain similar spectra, the components can be represented by either $\boldsymbol{F}$ or $\boldsymbol{H}$. This ambiguity will cause the following problem: part of the target source is captured by $\boldsymbol{HU}$ or part of the nontarget sources is mixed into $\boldsymbol{FG}$, which degrades the accuracy of source separation.

2.2 Conventional PSNMF Based on Inner Product

To solve the above-mentioned problem, PSNMF [9] was proposed. In PSNMF, the nontarget bases in $\boldsymbol{H}$ are forced to be as orthogonal as possible to the supervised bases in $\boldsymbol{F}$ by imposing an orthogonality penalty to the cost function of the separation stage as follows:

$$\begin{aligned}&\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{D}(\boldsymbol{Y}\|\boldsymbol{FG}+\boldsymbol{HU}) + \mu\mathcal{P}_{\mathrm{inner}}(\boldsymbol{F},\boldsymbol{H})\\ &\text{s.t.}\;\; g_{kj}, h_{il}, u_{lj} \geq 0 \;\; \forall i,\; j,\; k,\; l,\end{aligned} \tag{4}$$

where $\mu > 0$ is the weight coefficient and $\mathcal{P}_{\mathrm{inner}}(\boldsymbol{F},\boldsymbol{H})$ is defined as

$$\begin{aligned}\mathcal{P}_{\mathrm{inner}}(\boldsymbol{F},\boldsymbol{H}) &= \|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{Fr}}^2\\ &= \sum_{k,l}\left(\sum_i f_{ik}h_{il}\right)^2.\end{aligned} \tag{5}$$

The penalty term $\mathcal{P}_{\mathrm{inner}}(\boldsymbol{F},\boldsymbol{H})$ corresponds to the sum of squared inner products between two bases in $\boldsymbol{F}$ and $\boldsymbol{H}$.

Therefore, the nontarget basis matrix $\boldsymbol{H}$ is estimated by taking the following perspectives into account: (i) the divergence between $\boldsymbol{Y}$ and $\boldsymbol{FG}+\boldsymbol{HU}$ should be minimized and (ii) the bases in $\boldsymbol{H}$ should be as orthogonal as possible to the bases in $\boldsymbol{F}$. As a result, the source separation performance of PSNMF is greatly improved from that of SNMF [9].

The update rules for the variables $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ based on the MM algorithm [10] are respectively derived as follows [9]:

$$g_{kj} \leftarrow g_{kj} \cdot \frac{\sum_i \frac{y_{ij}}{\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}} f_{ik}}{\sum_i f_{ik}}, \tag{6}$$

$$h_{il} \leftarrow h_{il} \cdot \frac{\sum_j \frac{y_{ij}}{\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}} u_{lj}}{\sum_j u_{lj} + \mu f_{ik} \sum_{i'} f_{i'k} h_{i'l}}, \tag{7}$$

$$u_{lj} \leftarrow u_{lj} \cdot \frac{\sum_i \frac{y_{ij}}{\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}} h_{il}}{\sum_i h_{il}}, \tag{8}$$

where y_{ij} is the element of $\boldsymbol{Y}$.

3. Proposed Method

3.1 Motivation and Strategy

In conventional PSNMF Eq. (4), the penalty term $\mathcal{P}_{\text{inner}}(\boldsymbol{F},\boldsymbol{H})$ is minimized to increase the dissimilarity between $\boldsymbol{F}$ and $\boldsymbol{H}$. However, in practice, the penalty term $\mathcal{P}_{\text{inner}}(\boldsymbol{F},\boldsymbol{H})$ can be minimized by simply multiplying $\boldsymbol{H}$ by $\alpha > 0$ as

$$\begin{aligned}\lim_{\alpha\to+0} \mathcal{P}_{\text{inner}}(\boldsymbol{F},\alpha\boldsymbol{H}) &= \lim_{\alpha\to+0} \alpha\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{Fr}}^2 \\ &= 0. \end{aligned} \tag{9}$$

In this case, the activation matrix $\boldsymbol{U}$ can be updated as $\boldsymbol{U} \leftarrow \alpha^{-1}\boldsymbol{U}$ so that the value of divergence does not change, i.e.,

$$\mathcal{D}(\boldsymbol{Y}\|\boldsymbol{FG}+\alpha\boldsymbol{H}[\alpha^{-1}\boldsymbol{U}]) = \mathcal{D}(\boldsymbol{Y}\|\boldsymbol{FG}+\boldsymbol{HU}). \tag{10}$$

For this reason, the penalty term $\mathcal{P}_{\text{inner}}(\boldsymbol{F},\boldsymbol{H})$ does not affect the optimization of $\boldsymbol{H}$, and the minimization problem Eq. (4) is equivalent to Eq. (3). Therefore, the orthogonalization between the bases in $\boldsymbol{F}$ and $\boldsymbol{H}$ is not properly performed in conventional PSNMF Eq. (4).

To cope with this problem, in [9], PSNMF is implemented with a basis normalization process $\boldsymbol{h}_l \leftarrow \boldsymbol{h}_l/\|\boldsymbol{h}_l\|$ and $\boldsymbol{u}_l \leftarrow \|\boldsymbol{h}_l\|\boldsymbol{u}_l$ after every update of $\boldsymbol{H}$ so that the matrix $\boldsymbol{HU}$ does not change, where $\boldsymbol{h}_l = [h_{1l}, h_{2l}, \cdots, h_{Il}]^{\mathrm{T}}$, $\boldsymbol{u}_l = [u_{l1}, u_{l2}, \cdots, u_{lJ}]^{\mathrm{T}}$, and $\|\cdot\|$ is an arbitrary norm. Although the source separation performance is improved by this heuristic implementation technique, this iterative normalization process is not a fundamental (or mathematical) solution. In addition, this basis normalization can increase the value of cost function Eq. (4), hence theoretical convergence of the MM algorithm (the update rules Eqs. (6)–(8)) is lost.

A similar problem was pointed out in the context of sparse NMF [12], where sparse regularization based on the L_1 norm of the activation matrix is imposed to the cost function of simple NMF. For the sparse NMF, in [13], a norm constraint of bases is newly imposed to the NMF optimization to avoid the scale indeterminacy problem. Although PSNMF may be solved by introducing the same constraint as in [13], the theoretical convergence in this approach cannot be ensured. This is because the method [13] utilizes fraction-based update rules with positive and negative terms in the gradient. This heuristic update rule has often appeared in the history of NMF-based methods (e.g., [14], [15]), but its theoretical convergence with the norm constraint has not been proven.

In this paper, to solve the problem in conventional PSNMF, we propose to utilize cosine similarity as an orthogonality penalty term. Since cosine similarity depends on only the angle between two input vectors, we can measure the orthogonality regardless of the lengths of the vectors and the scale indeterminacy in NMF does not affect the penalty term. We propose two types of cosine similarity penalty: logarithmic cosine similarity (hereafter referred to as *Log-Cos PSNMF*) and simple cosine similarity (hereafter referred to as *Cos PSNMF*) penalties. In addition, we derive MM-algorithm-based (convergence-guaranteed) update rules for both Log-Cos PSNMF and Cos PSNMF and obtain a tuning-free and easy-to-use optimization algorithm.

3.2 Proposed Method 1: Log-Cos PSNMF

3.2.1 Cost Function

In Log-Cos PSNMF, we consider the following optimization problem:

$$\min_{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}} \mathcal{J}_1 \quad \text{s.t.} \quad g_{kj}, h_{il}, u_{lj} \geq 0 \ \forall i, j, k, l, \tag{11}$$

where

$$\mathcal{J}_1 = \mathcal{D}(\boldsymbol{Y}\|\boldsymbol{FG}+\boldsymbol{HU}) + \mu\mathcal{P}_{\text{logcos}}(\boldsymbol{F},\boldsymbol{H}), \tag{12}$$

$$\mathcal{P}_{\text{logcos}}(\boldsymbol{F},\boldsymbol{H}) = \sum_{k,l} \log \frac{\sum_i f_{ik} h_{il}}{\left(\sum_i f_{ik}^2\right)^{\frac{1}{2}} \left(\sum_i h_{il}^2\right)^{\frac{1}{2}}} \tag{13}$$

$$= \sum_{k,l} \left[\log \sum_i f_{ik} h_{il} - \frac{1}{2} \log \sum_i f_{ik}^2 - \frac{1}{2} \log \sum_i h_{il}^2 \right]. \tag{14}$$

The penalty term $\mathcal{P}_{\text{logcos}}(\boldsymbol{F},\boldsymbol{H})$ corresponds to the sum of logarithmic cosine similarities for all combinations of supervised bases in $\boldsymbol{F}$ and nontarget bases in $\boldsymbol{H}$. The logarithm function in Eq. (13) is employed to decompose the cosine similarity as Eq. (14), resulting in a simpler form of the MM-algorithm-based update rules compared with those of Cos PSNMF.

3.2.2 Derivation of MM-Algorithm-Based Update Rules

Since it is difficult to directly minimize the cost function Eq. (12), we use the MM algorithm [10] as in conventional PSNMF. Note that since the penalty term $\mathcal{P}_{\mathrm{logcos}}(\boldsymbol{F},\boldsymbol{H})$ in Eq. (12) does not depend on the activation matrices $\boldsymbol{G}$ and $\boldsymbol{U}$, the update rules for $\boldsymbol{G}$ and $\boldsymbol{U}$ that minimize Eq. (12) are equivalent to Eqs. (6) and (8), respectively.

The cost function Eq. (12) can be rewritten as follows:

$$\begin{aligned}\mathcal{J}_1 = \sum_{i,j}\Bigg[& y_{ij}\log y_{ij} - y_{ij}\log\left(\sum_k f_{ik}g_{kj} + \sum_l h_{il}u_{lj}\right)\\ &- y_{ij} + \sum_k f_{ik}g_{kj} + \sum_l h_{il}u_{lj}\Bigg]\\ &+ \mu\sum_{k,l}\Bigg[\log\sum_i f_{ik}h_{il} - \frac{1}{2}\log\sum_i f_{ik}^2\\ &- \frac{1}{2}\log\sum_i h_{il}^2\Bigg]. \end{aligned} \tag{15}$$

The second, sixth, and eighth terms of Eq. (15) contain the sum of the variables (g_{kj}, h_{il}, and u_{lj}) in a logarithm function, and it is difficult to calculate the stationary point w.r.t. the variables. In the MM algorithm, we design an upper bound function to indirectly optimize the cost function whose stationary points are difficult to calculate. Since the second term, i.e., the negative logarithmic function ($-\log(\cdot)$), is a convex function, by applying Jensen's inequality, we can design the upper bound function as follows:

$$\begin{aligned}&-\log\left(\sum_k f_{ik}g_{kj} + \sum_l h_{il}u_{lj}\right)\\ &\quad= -\log\left(\sum_k \frac{\alpha_{ijk}f_{ik}g_{kj}}{\alpha_{ijk}} + \sum_l \frac{\beta_{ijl}h_{il}u_{lj}}{\beta_{ijl}}\right)\\ &\quad\le -\sum_k \alpha_{ijk}\log\frac{f_{ik}g_{kj}}{\alpha_{ijk}} - \sum_l \beta_{ijl}\log\frac{h_{il}u_{lj}}{\beta_{ijl}}, \end{aligned} \tag{16}$$

where $\alpha_{ijk} > 0$ and $\beta_{ijl} > 0$ are auxiliary variables that satisfy $\sum_k \alpha_{ijk} + \sum_l \beta_{ijl} = 1$. Similarly to Eq. (16), we design the upper bound function of the eighth term of Eq. (15) as

$$\begin{aligned}-\log\sum_i h_{il}^2 &= -\log\sum_i \gamma_{il}\frac{h_{il}^2}{\gamma_{il}}\\ &\le -\sum_i \gamma_{il}\log\frac{h_{il}^2}{\gamma_{il}},\\ &= -2\sum_i \gamma_{il}\log h_{il} + \sum_i \gamma_{il}\log\gamma_{il}, \end{aligned} \tag{17}$$

where $\gamma_{il} > 0$ is an auxiliary variable that satisfies $\sum_i \gamma_{il} = 1$. The equality in Eqs. (16) and (17) respectively holds if and only if

$$\alpha_{ijk} = \frac{f_{ik}g_{kj}}{\sum_{k'} f_{ik'}g_{k'j} + \sum_l h_{il}u_{lj}}, \tag{18}$$

$$\beta_{ijl} = \frac{h_{il}u_{lj}}{\sum_k f_{ik}g_{kj} + \sum_{l'} h_{il'}u_{l'j}}, \tag{19}$$

$$\gamma_{il} = \frac{h_{il}^2}{\sum_{i'} h_{i'l}^2}. \tag{20}$$

The sixth term of Eq. (15) is a positive logarithm ($+\log(\cdot)$) function that includes the sum of the variables (h_{il}). Since the positive logarithmic function is a concave function, by applying the tangent-line inequality, the upper bound function can be designed as

$$\log\sum_i f_{ik}g_{kj} \le \frac{1}{\delta_{kli}}\left(\sum_i f_{ik}g_{kj} - \delta_{kli}\right) + \log\delta_{kli}, \tag{21}$$

where $\delta_{kli} > 0$ is an auxiliary variable. The equality in Eq. (21) holds if and only if

$$\delta_{kli} = \sum_i f_{ik}h_{il}. \tag{22}$$

From Eqs. (16), (17), and (21), the upper bound function $\mathcal{J}_1^+$ of $\mathcal{J}_1$ can be designed as

$$\begin{aligned}\mathcal{J}_1 &\le \mathcal{J}_1^+\\ &= \sum_{i,j}\Bigg[y_{ij}\log y_{ij}\\ &\quad - y_{ij}\left(\sum_k \alpha_{ijk}\log\frac{f_{ik}g_{kj}}{\alpha_{ijk}} + \sum_l \beta_{ijl}\log\frac{h_{il}u_{lj}}{\beta_{ijl}}\right)\\ &\quad - y_{ij} + \sum_k f_{ik}g_{kj} + \sum_l h_{il}u_{lj}\Bigg]\\ &\quad + \mu\sum_{k,l}\Bigg[\frac{1}{\delta_{kli}}\left(\sum_i f_{ik}h_{il} - \delta_{kli}\right) + \log\delta_{kli}\\ &\quad - \frac{1}{2}\log\sum_i f_{ik}^2 - \sum_i \gamma_{il}\log h_{il}\\ &\quad + \frac{1}{2}\sum_i \gamma_{il}\log\gamma_{il}\Bigg]. \end{aligned} \tag{23}$$

The update rule for h_{il} can be derived by solving $\partial\mathcal{J}_1^+/\partial h_{il} = 0$ w.r.t. h_{il} as

$$h_{il} = \frac{\sum_j y_{ij}\beta_{ijl} + \mu\sum_k \gamma_{il}}{\sum_j u_{lj} + \mu\sum_k \frac{f_{ik}}{\delta_{kli}}}. \tag{24}$$

By substituting the equality conditions Eqs. (19), (20), and (22) for Eq. (24), we obtain the multiplicative update rule for h_{il} as

$$h_{il} \leftarrow h_{il} \cdot \frac{\sum_j \frac{y_{ij}}{\sum_{k'} f_{ik'} g_{k'j} + \sum_{l'} h_{il'} u_{l'j}} u_{lj} + \mu K \frac{h_{il}}{\sum_{i'} h_{i'l}^2}}{\sum_j u_{lj} + \mu \sum_k \frac{f_{ik}}{\sum_{i'} f_{i'k} h_{i'l}}}. \tag{25}$$

Note that Eq. (25) coincides with Eq. (7) when $\mu = 0$.

By iterating the update rule Eq. (25), we can estimate $\boldsymbol{H}$ that tends to be orthogonal to $\boldsymbol{F}$. However, in Log-Cos PSNMF, the penalty term $\mathcal{P}_{\text{logcos}}(\boldsymbol{F}, \boldsymbol{H})$ becomes $-\infty$ when $\boldsymbol{h}_l$ is perfectly orthogonal to any bases in $\boldsymbol{F}$ or $\boldsymbol{h}_l = \boldsymbol{0}$ for any l. In such a case, the update rule Eq. (25) is undefined, and the solution of minimization problem Eq. (11), $\mathcal{J}_1 \rightarrow -\infty$, is meaningless. To avoid this inherent problem, the following flooring process [16] is applied in each iteration:

$$h_{il} \leftarrow \max(h_{il}, \varepsilon), \tag{26}$$

where ε is the machine epsilon and $\max(\cdot)$ is a function that returns the largest value of the input arguments.

3.3 Proposed Method 2: Cos PSNMF

3.3.1 Cost Function

Log-Cos PSNMF utilizes a logarithm function in the penalty $\mathcal{P}_{\text{logcos}}(\boldsymbol{F}, \boldsymbol{H})$ to decompose the cosine similarity function. This logarithm function makes the cost function $\mathcal{J}_1$ unbounded, namely, the minimum value of $\mathcal{J}_1$ becomes $-\infty$. We apply the flooring Eq. (26) to avoid the meaningless solution, but such heuristic treatment is not essential as well as the basis normalization described in Sect. 3.1. To cope with this problem, we propose Cos PSNMF, which utilizes a simple cosine similarity penalty. Although derivation and update rules become more complicated compared with those in Log-Cos PSNMF, we can still obtain a convergence-guaranteed optimization algorithm in this formulation.

We consider the following optimization problem:

$$\min_{\boldsymbol{G}, \boldsymbol{H}, \boldsymbol{U}} \mathcal{J}_2 \quad \text{s.t.} \quad g_{kj}, h_{il}, u_{lj} \geq 0 \ \ \forall i, j, k, l, \tag{27}$$

where

$$\mathcal{J}_2 = \mathcal{D}(\boldsymbol{Y} \| \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}) + \mu \mathcal{P}_{\cos}(\boldsymbol{F}, \boldsymbol{H}), \tag{28}$$

$$\mathcal{P}_{\cos}(\boldsymbol{F}, \boldsymbol{H}) = \sum_{k,l} \frac{\sum_i f_{ik} h_{il}}{\left(\sum_i f_{ik}^2\right)^{\frac{1}{2}} \left(\sum_i h_{il}^2\right)^{\frac{1}{2}}}. \tag{29}$$

In Cos PSNMF, the logarithm function in $\mathcal{P}_{\text{logcos}}(\boldsymbol{F}, \boldsymbol{H})$ is omitted, and the minimum value of the cost function Eq. (28) is bounded by zero.

3.3.2 Derivation of MM-Algorithm-Based Update Rules

Similarly to Log-Cos PSNMF, direct minimization of the cost function Eq. (27) is difficult, and we apply the MM algorithm. The update rules of $\boldsymbol{G}$ and $\boldsymbol{U}$ are equivalent to Eqs. (6) and (8), respectively.

The penalty term $\mathcal{P}_{\cos}(\boldsymbol{F}, \boldsymbol{H})$ includes $(\sum_i h_{il}^2)^{-1/2}$, and this function hinders the direct calculation of a stationary point w.r.t. h_{il}. Since this function is convex, we can design the upper bound function by applying Jensen's inequality as follows:

$$\begin{aligned}\left(\sum_i h_{il}^2\right)^{-\frac{1}{2}} &= \left(\sum_i \epsilon_{il} \frac{h_{il}^2}{\epsilon_{il}}\right)^{-\frac{1}{2}} \\ &\leq \sum_i \epsilon_{il} \left(\frac{h_{il}^2}{\epsilon_{il}}\right)^{-\frac{1}{2}} \\ &= \sum_i \epsilon_{il}^{\frac{3}{2}} h_{il}^{-1},\end{aligned} \tag{30}$$

where $\epsilon_{il} > 0$ is an auxiliary variable that satisfies $\sum_i \epsilon_{il} = 1$. The equality in Eq. (30) holds if and only if

$$\epsilon_{il} = \frac{h_{il}^2}{\sum_{i'} h_{i'l}^2}. \tag{31}$$

From Eqs. (16) and (30), the upper bound function $\mathcal{J}_2^+$ of $\mathcal{J}_2$ can be designed as

$$\begin{aligned}\mathcal{J}_2 &\leq \mathcal{J}_2^+ \\ &= \sum_{i,j} \Bigg[y_{ij} \log y_{ij} \\ &\quad - y_{ij} \left(\sum_k \alpha_{ijk} \log \frac{f_{ik} g_{kj}}{\alpha_{ijk}} + \sum_l \beta_{ijl} \log \frac{h_{il} u_{lj}}{\beta_{ijl}} \right) \\ &\quad - y_{ij} + \sum_k f_{ik} g_{kj} + \sum_l h_{il} u_{lj} \Bigg] \\ &\quad + \mu \sum_{k,l} \left[\left(\sum_{i'} f_{i'k}^2 \right)^{-\frac{1}{2}} \left(\sum_{i',i''} f_{i'k} \epsilon_{i''l}^{\frac{3}{2}} \frac{h_{i'l}}{h_{i''l}} \right) \right].\end{aligned} \tag{32}$$

The update rule for h_{il} can be derived by solving $\partial \mathcal{J}_2^+ / \partial h_{il} = 0$ w.r.t. h_{il}. The derivative can be obtained as

$$\begin{aligned}&\sum_j \left[\frac{-y_{ij} \beta_{ijl}}{h_{il}} + u_{lj} \right] + \mu \sum_k \left[\left(\sum_{i'} f_{i'k}^2 \right)^{-\frac{1}{2}} \right. \\ &\left. \left(-\frac{1}{h_{il}^2} \epsilon_{il}^{\frac{3}{2}} \sum_{i' \neq i} f_{i'k} h_{i'l} + f_{ik} \sum_{i' \neq i} \epsilon_{i'l}^{\frac{3}{2}} \frac{1}{h_{i'l}} \right) \right] = 0.\end{aligned} \tag{33}$$

Rearrangement of Eq. (33) gives the quadratic equation

$$a_{il} h_{il}^2 + b_{il} h_{il} + c_{il} = 0, \tag{34}$$

where

$$a_{il} = \sum_j u_{lj} + \mu \sum_k f_{ik} \left(\sum_{i'} f_{i'k}^2 \right)^{-\frac{1}{2}} \left(\sum_{i' \neq i} \epsilon_{i'l}^{\frac{3}{2}} \frac{1}{h_{i'l}} \right), \tag{35}$$

$$b_{il} = -\sum_j y_{ij} \beta_{ijl}, \tag{36}$$

$$c_{il} = -\mu\epsilon_{il}^{\frac{3}{2}} \sum_{k} \left(\sum_{i'} f_{i'k}^{2}\right)^{-\frac{1}{2}} \left(\sum_{i' \neq i} f_{i'k} h_{i'l}\right). \tag{37}$$

Thus, the stationary point of $\mathcal{J}_2^{+}$ w.r.t. h_{il} is given by the quadratic formula:

$$h_{il} = \frac{-b_{il} \pm \sqrt{b_{il}^{2} - 4a_{il}c_{il}}}{2a_{il}}, \tag{38}$$

where ± in Eq. (38) should be determined so that the r.h.s. of Eq. (38) becomes nonnegative.

Similarly to Eq. (25), a unified update rule of h_{il} can be obtained by substituting the equality conditions Eqs. (19), (20), and (31) for Eq. (38). Since the unified update rule becomes complicated, we instead show the update rules of a_{il}, b_{il}, and c_{il}:

$$a_{il} \leftarrow \sum_{j} u_{lj} + \mu \sum_{k} f_{ik} \left(\sum_{i'} f_{i'k}^{2}\right)^{-\frac{1}{2}} \left[\sum_{i' \neq i} h_{i'l}^{2} \left(\sum_{i''} h_{i''l}^{2}\right)^{-\frac{3}{2}}\right], \tag{39}$$

$$b_{il} \leftarrow -\sum_{j} y_{ij} \frac{h_{il}u_{lj}}{\sum_{k} f_{ik}g_{kj} + \sum_{l'} h_{il'}u_{l'j}}, \tag{40}$$

$$c_{il} \leftarrow -\mu \left(\frac{h_{il}^{2}}{\sum_{i'} h_{i'l}^{2}}\right)^{\frac{3}{2}} \sum_{k} \left(\sum_{i'} f_{i'k}^{2}\right)^{-\frac{1}{2}} \left(\sum_{i' \neq i} f_{i'k} h_{i'l}\right). \tag{41}$$

The up-to-date variable h_{il} is obtained by calculating Eq. (38) after the update of Eqs. (39)–(41). Note that Eqs. (38)–(41) coincide with Eq. (7) when $\mu = 0$.

4. Experiment

4.1 Conditions

To confirm the validity of the proposed methods, we compare the performances of simple SNMF, conventional PSNMF [9] (*Inner-Prod. PSNMF*), Log-Cos PSNMF, and Cos PSNMF. In the experiments, we used the artificial audio sources produced by a Yamaha MU-1000 synthesizer in songKitamura [6], [17] as the development and test datasets. Two instrumental melodies were selected from the eleven instruments, namely, oboe, trumpet, horn, flute, violin, clarinet, piano, harpsichord, trombone, bassoon, and cello, and we mixed them with the same power to produce a two-source mixture monaural signal $\boldsymbol{Y}$, resulting in 90 mixture signals. Then, we randomly split 90 mixture signals into 45 development dataset and 45 test dataset. The development dataset was used to find the optimal hyperparameter μ in each method, and the test dataset was used for performance comparison. For the sample sound in the training stage, two-octave ascending notes of the target instrument were used as $\boldsymbol{X}$ for obtaining the supervised basis matrix $\boldsymbol{F}$.

The source-to-distortion ratio (SDR) [18] was used as the evaluation score, which includes both the quality of the separated target sound (absence of artificial noise) and the degree of separation (absence of nontarget source components). The initial values of each matrix were set to uniformly distributed random values in (0, 1). The window and shift lengths in short-time Fourier transform were set to 92.9 ms and 46.4 ms, respectively. The numbers of bases for the target and nontarget sources were $K = 27$ (24 notes and three common spectra) and $L = 50$, respectively.

For Inner-Prod. PSNMF, as described in Sect. 3.1, we must apply the following basis normalization to validate the penalty term $\mathcal{P}_{\mathrm{inner}}(\boldsymbol{F}, \boldsymbol{H})$:

$$h_{il} \leftarrow \frac{1}{d_l} h_{il}, \tag{42}$$

$$u_{lj} \leftarrow d_l u_{lj}, \tag{43}$$

where $d_l = \sum_i h_{il}$. This normalization was performed after the update of all the parameters in each iteration. Also, for Log-Cos PSNMF, we performed both Eq. (26) (after the update of h_{il}) and Eqs. (42) and (43) to avoid numerical instability.

4.2 Results

4.2.1 Hyperparameter Tuning Using Development Dataset

Figure 3 shows average SDR behaviors for the development dataset in each method. We can confirm that two proposed methods, Log-Cos PSNMF and Cos PSNMF, outperform Inner-Prod. PSNMF at the optimal hyperparameter setting. This is because the penalty terms $\mathcal{P}_{\mathrm{logcos}}(\boldsymbol{F}, \boldsymbol{H})$ and $\mathcal{P}_{\mathrm{cos}}(\boldsymbol{F}, \boldsymbol{H})$ appropriately force the orthogonality between the supervised and nontarget bases in the optimization of $\boldsymbol{H}$, whereas the conventional penalty term $\mathcal{P}_{\mathrm{inner}}(\boldsymbol{F}, \boldsymbol{H})$ does not directly affect the optimization because of the scale indeterminacy in NMF; the scale of $\boldsymbol{H}$ simply decreases when we set a large μ value. However, since the proposed methods

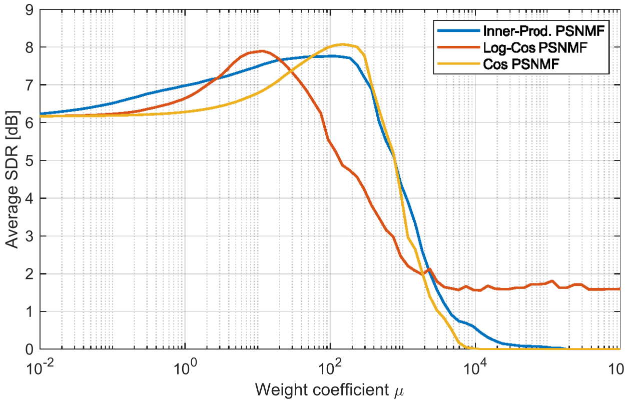

Fig. 3 Average SDR behaviors of development dataset. Weight coefficient μ that provides highest average SDR in development dataset is used for performance comparison with test dataset.

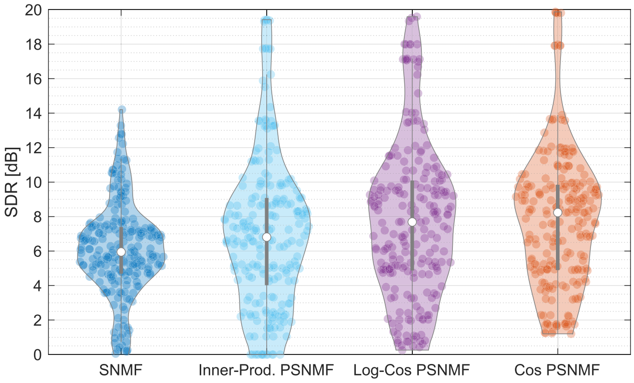

Fig. 4 Violin plot of 45 SDR results in test dataset. In each method, white circle indicates median value, gray vertical line shows range of 25–75 percentiles, and violin-shape curve is estimated distributions.

Table 1 Average and median values for test dataset.

Method	Average [dB]	Median [dB]
SNMF	6.07	5.95
Inner-Prod. PSNMF	7.01	6.81
Log-Cos PSNMF	7.75	7.69
Cos PSNMF	**7.82**	**8.23**

eliminate the scale indeterminacy problem, the optimal μ becomes relatively peaky compared with that of the conventional PSNMF.

4.2.2 SDR Comparison Using Test Dataset

The SDR values of each method for the test dataset were compared, where the hyperparameter μ in each PSNMF was set to the optimal value that provides the highest scores in the development dataset (Fig. 3). The violin plot of 45 SDR results is shown in Fig. 4. Also, average and median values of the results are summarized in Table 1. We can confirm that both Log-Cos PSNMF and Cos PSNMF outperform the Inner-Prod. PSNMF in terms of average and median values. In particular, Cos PSNMF provides the best median value, whose improvement from that of Inner-Prod. PSNMF is 1.42 dB. This result shows that the optimal settings of the hyperparameter μ in Log-Cos PSNMF and Cos PSNMF do not strongly depend on the type of instruments.

4.2.3 Statistical Testing

To evaluate the validity of SDR differences in Table 1, two types of statistical testing were applied, namely, the pairwise one-sided Welch's t test [19] and the pairwise Brunner–Munzel (BM) test [20]. The null hypotheses of the one-sided Welch's t test and the BM test are respectively as follows:

- The true means μ_{A} and μ_{B} of normally distributed samples in two groups A and B satisfy $\mu_{\mathrm{A}} \geq \mu_{\mathrm{B}}$, where variances of them are unequal.
- When we randomly select samples s_A and s_B from each of two groups A and B, the probabilities of $s_A \geq s_B$ and $s_A < s_B$ are equal (stochastic equality).

Table 2 p values obtained by pairwise one-sided Welch's t test.

Method of group A	Method of group B	p value [%]
SNMF	Inner-Prod. PSNMF	0.3389
SNMF	Log-Cos PSNMF	0.0001
SNMF	Cos PSNMF	0.0000
Inner-Prod. PSNMF	Log-Cos PSNMF	3.9434
Inner-Prod. PSNMF	Cos PSNMF	2.0614
Log-Cos PSNMF	Cos PSNMF	42.8965

Table 3 p values obtained by pairwise BM test.

Method of group A	Method of group B	p value [%]
SNMF	Inner-Prod. PSNMF	3.1576
SNMF	Log-Cos PSNMF	0.0056
SNMF	Cos PSNMF	0.0000
Inner-Prod. PSNMF	Log-Cos PSNMF	5.9094
Inner-Prod. PSNMF	Cos PSNMF	0.9811
Log-Cos PSNMF	Cos PSNMF	59.7709

Note that BM test does not assume the normal distribution for observed data.

The p values obtained by each test are shown in Tables 2 and 3, respectively. When we compare the conventional and proposed methods, the null hypotheses in each test can be rejected with satisfactory confidence. For example, the p values for comparing Inner-Prod. PSNMF and Cos PSNMF are 2.06% and 0.98% in one-sided Welch's t test and BM test, respectively, which are satisfactory low. These results show the efficacy of the proposed methods.

5. Conclusion

In this paper, we proposed new algorithms for SNMF using the cosine similarity penalty, where the scale indeterminacy in NMF does not affect the optimization of basis and activation matrices. From the results of experiments using mixtures of instrumental sources, we confirmed that the proposed PSNMF using the cosine-similarity-based orthogonality penalty can improve the separation performance compared with the conventional PSNMF using the inner-product-based orthogonality penalty. Since the parameter tuning for the proposed PSNMF becomes peaky, the prediction of the optimal μ is our important future work.

Although we only focus on SNMF in this paper, it is worth mentioning that the cosine similarity penalty can be applied even in the simple (unsupervised) NMF decomposition Eq. (1) for all the bases in $\boldsymbol{F}$. Since the orthogonality between all the bases can be maximized, NMF with cosine similarity can be used for a discriminative basis learning [21] and as an alternative approach to sparse NMF.

Acknowledgments

This work was partially supported by JSPS Grants-in-Aid for Scientific Research 19K20306 and 19H01116.

References

[1] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol.401, no.6755, pp.788–791, 1999.

[2] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," Proc. Neural Information Processing Systems, pp.556–562, 2000.

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

[4] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," Proc. WASPAA, pp.121–124, 2009.

[5] H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," Proc. ICASSP, pp.5365–5368, 2012.

[6] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," IEEE/ACM Trans. Audio, Speech, Language Process., vol.23, no.4, pp.654–669, 2015.

[7] 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. Audio, Speech, Language Process., vol.24, no.9, pp.1626–1641, 2016.

[8] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," Proc. LVA/ICA, pp.414–421, 2007.

[9] 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, vol.E97-A, no.5, pp.1113–1118, May 2014.

[10] D.R. Hunter and K. Lange, "Quantile regression via an MM algorithm," J. Comput. Graph. Stat., vol.9, no.1, pp.60–77, 2000.

[11] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," Proc. Irish Signals Syst. Conf., 2009.

[12] P.D. O'Grady and B.A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," Neurocomputing, vol.72, no.1, pp.88–101, 2008.

[13] J. Le Roux, F. Weninger, and J.R. Hershey, "Sparse NMF - half-baked or well done?," Mitsubishi Electric Research Lab. Technical Report, TR2015-023, 2015.

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

[15] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," Proc. WASPAA, 2015.

[16] N. Takahashi and M. Seki, "Multiplicative update for a class of constrained optimization problems related to NMF and its global convergence," Proc. EUSIPCO, pp.438–442, 2016.

[17] D. Kitamura, "Open dataset: songKitamura" http://d-kitamura.net/dataset_en.html, accessed Oct. 8. 2020.

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

[19] B.L. Welch, "The generalization of 'Student's' problem when several different population variances are involved," Biometrika, vol.34, no.1/2, pp.28–35, 1947.

[20] E. Brunner and U. Munzel, "The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation," Biometrical J., vol.42, no.1, pp.17–25, 2000.

[21] H. Nakajima, D. Kitamura, N. Takamune, H. Saruwatari, and N. Ono, "Bilevel optimization using stationary point of lower-level objective function for discriminative basis learning in nonnegative matrix factorization," IEEE Signal Process. Lett., vol.26, no.6, pp.818–822, 2019.

Yuta Iwase is a bachelor student of Advanced Course in Electronics, Information and Communication Engineering, National Institute of Technology, Kagawa College. His research interests include nonnegative matrix factorization and audio signal processing.

Daichi Kitamura received the Ph.D. degree from SOKENDAI, Hayama, Japan. He joined The University of Tokyo in 2017 as a Research Associate, and he moved to National Institute of Technology, Kagawa Collage as an Assistant Professor in 2018. His research interests include audio source separation, statistical signal processing, and machine learning. He was the recipient of the Awaya Prize Young Researcher Award from The Acoustical Society of Japan (ASJ) in 2015, Ikushi Prize from Japan Society for the Promotion of Science in 2017, Best Paper Award from IEEE Signal Processing Society Japan in 2017, Itakura Prize Innovative Young Researcher Award from ASJ in 2018, and IEEE Signal Processing Society (SPS) 2019 Young Author Best Paper Award from IEEE SPS in 2020.