Experimental Evaluation of Superresolution-Based Nonnegative Matrix Factorization for Binaural Recording

Daichi Kitamura[1,a)] Hiroshi Saruwatari[1] Satoshi Nakamura[1] Yu Takahashi[2]
Kazunobu Kondo[2] Hirokazu Kameoka[3]

Abstract: In this paper, we propose a new stereo signal separation scheme based on multi-divergence supervised nonnegative matrix factorization (SNMF). In previous studies, a hybrid method, which concatenates superresolution-based SNMF after directional clustering, has been proposed for multichannel signal separation. However, the optimal divergence in SNMF temporally fluctuates because the separation and extrapolation abilities depend on spatial conditions of sources in music tunes. To solve this problem, we propose a new scheme for multi-divergence, where optimal divergence can be automatically changed in each time frame according to the local spatial conditions. Experimental results show the proposed method efficacy.

1. Introduction

In recent years, music and acoustic signal separation based on nonnegative matrix factorization (NMF) [1] has been a very active area of signal-processing research [2], [3], [4], [5]. NMF of acoustic signals decomposes an input spectrogram into the product of a spectral basis matrix and its activation matrix. In particular, supervised NMF (SNMF) [6], [7], [8], [9], which includes a priori training with some sample sounds of a target instrument, can extract the target signal to some extent. However, in the case of a mixture consisting of many sources, the source extraction performance is markedly degraded when only single-channel observation is available.

Multichannel NMF, which is a natural extension of NMF to a multichannel signal, has been proposed as an unsupervised method [10], [11]. However, such an unsupervised separation is a difficult problem because the decomposition is underspecified. Hence, algorithms used for multichannel NMF have strong dependence on initial values and lack robustness.

As another means of addressing multichannel signal separation, a hybrid method, which concatenates superresolution-based SNMF after directional clustering, has been proposed by the authors [12]. This method uses index information generated by binary masking of directional clustering so that spectral chasms can be regarded as *unseen* observations, and finally reconstructs the target source components via spectrogram extrapolation using the supervised bases as a *dictionary*. Also, we have proposed some update algorithms for superresolution-based SNMF based on generalized Kullback-Leibler divergence (*KL-divergence*) and Euclidean distance (*EUC-distance*) [12]. In general SNMF-based music signal separation, KL-divergence is often used as a cost function because the spectrogram of music signals tends to become sparse, and KL-divergence-based SNMF is suitable for representing such sparse signals [13]. However, it has been experimentally confirmed that the optimal divergence for superresolution-based SNMF is EUC-distance [12]. This performance difference between EUC-distance and KL-divergence is due to the different spatial condition of sources, and the optimal divergence temporally fluctuates because the spatial condition is not consistent in general music tunes.

To solve this problem, we propose a new scheme for multi-divergence, where optimal divergence can be automatically changed in each time frame according to the local spatial conditions. This method can achieve the highest separation accuracy in various types of stereo signals. Experimental results in artificial and real-recorded signals confirm that the proposed method outperforms the single-divergence methods.

2. Conventional Method

2.1 SNMF

In SNMF, a priori spectral patterns (bases) should be trained in advance as a basis dictionary. The following equation represents the decomposition in SNMF:

$$\boldsymbol{Y} \simeq \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}, \qquad (1)$$

where $\boldsymbol{Y}(\in \mathbb{R}_{\geq 0}^{\Omega\times T})$ is an observed spectrogram, $\boldsymbol{F}(\in \mathbb{R}_{\geq 0}^{\Omega\times K})$ is a supervised basis matrix trained in advance, which includes spectral patterns of the target signal as column vectors, $\boldsymbol{G}(\in \mathbb{R}_{\geq 0}^{K\times T})$ is the activation matrix that corresponds to $\boldsymbol{F}$, $\boldsymbol{H}(\in \mathbb{R}_{\geq 0}^{\Omega\times L})$ represents the residual spectral patterns that cannot be expressed by $\boldsymbol{F}\boldsymbol{G}$, and $\boldsymbol{U}(\in \mathbb{R}_{\geq 0}^{L\times T})$ is the activation matrix that corresponds to $\boldsymbol{H}$. Moreover, Ω is the number of frequency bins, T is the number of frames of the observed signal, K is the number of bases of $\boldsymbol{F}$,

[1] Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan.
[2] Yamaha Corporation, Iwata, Shizuoka 438–0192, Japan.
[3] The University of Tokyo, Bunkyo-ku, Tokyo 113–8656, Japan.
[a)] daichi-k@is.naist.jp

and L is the number of bases of $\boldsymbol{H}$. In SNMF, the matrices $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ are optimized under the condition that $\boldsymbol{F}$ is known in advance. The matrix $\boldsymbol{F}$ can be trained by solving $\hat{\boldsymbol{Y}} = \boldsymbol{F}\hat{\boldsymbol{G}}$, where $\hat{\boldsymbol{Y}}$ is a training data spectrogram and $\hat{\boldsymbol{G}}$ is the corresponding activation matrix. Hence, $\boldsymbol{FG}$ ideally represents the target instrumental components and $\boldsymbol{HU}$ represents other components different from the target sounds after the decomposition.

SNMF can extract the target signal, particularly in the case of a small number of sources. However, in the case of a mixture consisting of many sources, the source extraction performance is markedly degraded because of the existence of similar-timbre instruments.

2.2 Directional Clustering and Its Hybrid Method Using Superresolution-Based SNMF

Decomposition methods employing directional information for the multichannel signal have also been proposed as unsupervised techniques [14], [15]. These methods quantize the direction via time-frequency binary masking. Such directional clustering works well, even in an underdetermined situation. However, there is an inherent problem that the sources located in the same direction cannot be separated only using directional information. Furthermore, the separated signal is likely to be distorted because the signal has many spectral chasms resulting from the binary-masking procedure as shown in Fig. 1.

To solve this problem, a hybrid method that concatenates superresolution-based SNMF after directional clustering has been proposed [12]. This SNMF algorithm explicitly utilizes index information determined by time-frequency binary masking in directional clustering. For example, if the target instrument is localized in the center cluster along with the interference, superresolution-based SNMF is only applied to the existing center components using index information (see Fig. 1). Therefore, the spectrogram of the target instrument is reconstructed using better matched bases because spectral chasms are treated as *unseen*, and these chasms have no impact on the cost function in SNMF. In addition, the components of the target instrument lost after directional clustering can be extrapolated using the supervised bases. In other words, the resolution of the target spectrogram is recovered with the superresolution by the supervised basis extrapolation. Furthermore, a regularization term is added in the cost function to avoid a basis extrapolation error [16].

2.3 Cost Function for Superresolution-Based SNMF

Here, the index matrix $\boldsymbol{I}(\in \mathbb{R}_{\{0,1\}}^{\Omega\times T})$ is obtained from the binary masking preceding the directional clustering. This index matrix has specific entries of unity or zero, which indicate whether or not each grid of the spectrogram belongs to the target directional cluster. The cost function in superresolution-based SNMF is defined using the index matrix $\boldsymbol{I}$ as [12]

$$\mathcal{J} = \sum_{\omega,t} i_{\omega,t}\mathcal{D}_{\beta_{\mathrm{NMF}}}\left(y_{\omega,t}\|\sum_k f_{\omega,k}g_{k,t} + \sum_l h_{\omega,l}u_{l,t}\right) + \lambda\sum_{\omega,t}\overline{i_{\omega,t}}\mathcal{D}_{\beta_{\mathrm{reg}}}\left(0\|\sum_k f_{\omega,k}g_{k,t}\right) + \mu\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{Fr}}^2, \quad (2)$$

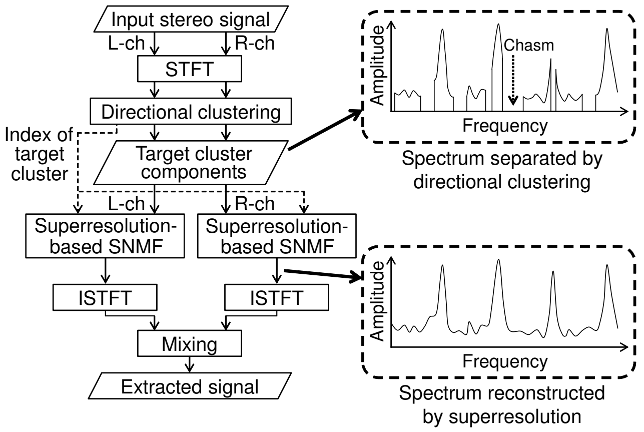
Input stereo signal
L-ch
R-ch
STFT
Directional clustering
Index of target cluster
Target cluster components
L-ch
R-ch
Superresolution-based SNMF
Superresolution-based SNMF
ISTFT
ISTFT
Mixing
Extracted signal
Amplitude
Chasm
Frequency
Spectrum separated by directional clustering
Amplitude
Frequency
Spectrum reconstructed by superresolution

Fig. 1 Signal flow of conventional hybrid method.

where $i_{\omega,t}$, $y_{\omega,t}$, $f_{\omega,k}$, $g_{k,t}$, $h_{\omega,l}$, and $u_{l,t}$ are the nonnegative entries of matrices $\boldsymbol{I}$, $\boldsymbol{Y}$, $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$, respectively, λ and μ are the weighting parameters for each penalty term, $\bar{\cdot}$ represents the binary complement of each entry in the index matrix, and $\circ$ indicates the Hadamard product of matrices. In addition, $\mathcal{D}_\beta(\cdot\|\cdot)$ is β-divergence, which is defined as [17]

$$\mathcal{D}_\beta(y\|x) = \begin{cases} \dfrac{y^\beta}{\beta(\beta-1)} + \dfrac{x^\beta}{\beta} - \dfrac{yx^{\beta-1}}{(\beta-1)} & (\beta\in\mathbb{R}_{\setminus\{0,1\}}) \\ y(\log y - \log x) + x - y & (\beta=1) \\ \dfrac{y}{x} - \log\dfrac{y}{x} - 1 & (\beta=0) \end{cases}. \quad (3)$$

This generalized divergence is a family of cost functions parameterized by a single shape parameter β that takes Itakura-Saito divergence, KL-divergence, and EUC-distance in special cases ($\beta = 0, 1$, and 2, respectively).

The update rules of superresolution-based SNMF based on KL-divergence are given by [12]

$$g_{k,t} \leftarrow \frac{g_{k,t}\sum_\omega i_{\omega,t}y_{\omega,t}f_{\omega,k}s_{\omega,t}^{-1}}{\sum_\omega i_{\omega,t}f_{\omega,k} + \lambda\sum_\omega \overline{i_{\omega,t}}f_{\omega,k}\sum_{k'} f_{\omega,k'}g_{k',t}}, \quad (4)$$

$$h_{\omega,l} \leftarrow \frac{h_{\omega,l}\sum_t i_{\omega,t}y_{\omega,t}u_{l,t}s_{\omega,t}^{-1}}{\sum_t i_{\omega,t}u_{l,t} + \mu\sum_k f_{\omega,k}\sum_{\omega'} f_{\omega',k}h_{\omega',l}}, \quad (5)$$

$$u_{l,t} \leftarrow \frac{u_{l,t}\sum_\omega i_{\omega,t}y_{\omega,t}h_{\omega,l}s_{\omega,t}^{-1}}{\sum_\omega i_{\omega,t}h_{\omega,l}}, \quad (6)$$

where

$$s_{\omega,t} = \sum_{k'} f_{\omega,k'}g_{k',t} + \sum_{l'} h_{\omega,l'}u_{l',t}. \quad (7)$$

Also, the update rules of superresolution-based SNMF based on EUC-distance are given by [12]

$$g_{k,t} \leftarrow \frac{g_{k,t}\sum_\omega i_{\omega,t}y_{\omega,t}f_{\omega,k}}{\sum_\omega i_{\omega,t}f_{\omega,k}s_{\omega,t} + \lambda\sum_\omega \overline{i_{\omega,t}}f_{\omega,k}\sum_{k'} f_{\omega,k'}g_{k',t}}, \quad (8)$$

$$h_{\omega,l} \leftarrow \frac{h_{\omega,l}\sum_t i_{\omega,t}y_{\omega,t}u_{l,t}}{\sum_t i_{\omega,t}u_{l,t}s_{\omega,t} + \mu\sum_k f_{\omega,k}\sum_{\omega'} f_{\omega',k}h_{\omega',l}}, \quad (9)$$

$$u_{l,t} \leftarrow \frac{u_{l,t}\sum_\omega i_{\omega,t}y_{\omega,t}h_{\omega,l}}{\sum_\omega i_{\omega,t}h_{\omega,l}s_{\omega,t}}. \quad (10)$$

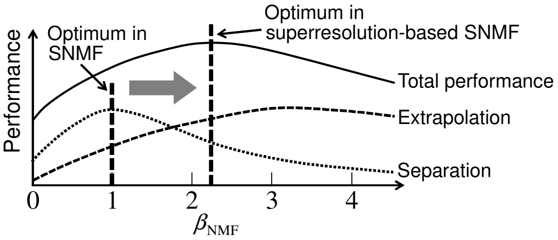

Fig. 2 Trade-off between separation and extrapolation.

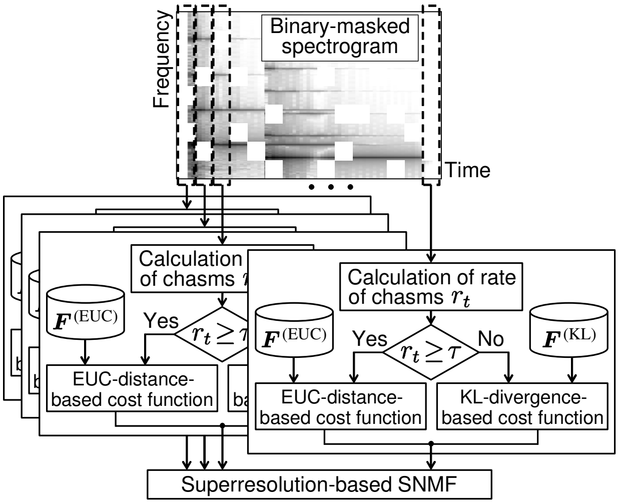

Fig. 3 Multi-divergence algorithm of proposed method.

3. Proposed Method

3.1 Divergence Dependence on Local Chasm Condition

In general SNMF-based music signal separation, KL-divergence is often used as a cost function because KL-divergence-based SNMF decomposes the observed spectrogram into the mixture of sparser bases than EUC-distance-based SNMF. The spectrogram of music signals tends to become sparse, and KL-divergence-based SNMF is suitable for representing such sparse signals [7], [13]. However, it has been experimentally confirmed that the optimal divergence for superresolution-based SNMF is EUC-distance [12]. This discrepancy in the divergence is due to the fact that superresolution-based SNMF has two tasks, namely, *signal separation* and *basis extrapolation*. The sparseness is not suitable for basis extrapolation because it is difficult to extrapolate sparse bases only from observable data. Figure 2 shows the trade-off between of the sparseness and separation performance of superresolution-based SNMF. A sparse criterion, such as KL-divergence ($\beta_{\mathrm{NMF}} = 1$), is not suitable for superresolution-based SNMF.

The optimal divergence for superresolution-based SNMF depends on the rate of spectral chasms in each time frame of the spectrogram obtained by directional clustering because of the trade-off between the separation and extrapolation abilities. If there are many chasms in a frame of a binary-masked spectrogram, high extrapolation ability is preferable for superresolution-based SNMF. In contrast, if the rate of chasms is low, separation ability is required rather than extrapolation ability. Therefore, it is expected that EUC-distance should be used in frames that have many chasms and KL-divergence should be used in the other frames. To improve the separation performance of superresolution-based SNMF for all types of input signals, we propose a new multi-divergence method as described below.

3.2 Cost Function Based on Multi-Divergence

Considering the above-mentioned divergence dependence on the local chasm condition, we propose to adapt the divergence in each frame of the spectrogram to the optimal one according to the rate of chasms in each frame r_t and a threshold value τ ($0 \leq \tau \leq 1$), where the rate of chasms r_t can be calculated from the index matrix $\boldsymbol{I}$. Straightforward but naive extension to this purpose is to apply independent SNMF to the short time-period data with switching the divergence in an online manner (hereafter referred to as *online hybrid method*). In this method, however, the size of each input matrix becomes small and the dimensionality is reduced. This degrades the separation performance because the trained bases $\boldsymbol{F}$ can represent any small-dimension matrix and no component is pushed into the interference $\boldsymbol{HU}$.

To cope with the problem and maintain the sufficient dimensionality of the matrix, we propose a new batch SNMF with a multi-divergence-based cost function covered onto the whole input matrix (see Fig. 3). The proposed cost function is defined as

$$\mathcal{J} = \sum_t \mathcal{J}_t, \tag{11}$$

$$\mathcal{J}_t = \begin{cases} \sum_\omega i_{\omega,t}\mathcal{D}_{\beta=2}(y_{\omega,t}\|s^{(\mathrm{EUC})}_{\omega,t}) \\ \quad + \lambda \sum_\omega \overline{i_{\omega,t}}\mathcal{D}_{\beta_{\mathrm{reg}}}(0\|\sum_k f^{(\mathrm{EUC})}_{\omega,k} g_{k,t}) \\ \quad + \mu\|\boldsymbol{F}^{(\mathrm{EUC})\mathrm{T}}\boldsymbol{H}\|^2_{\mathrm{Fr}} \qquad (r_t \geq \tau) \\ \sum_\omega i_{\omega,t}\mathcal{D}_{\beta=1}(y_{\omega,t}\|s^{(\mathrm{KL})}_{\omega,t}) \\ \quad + \lambda \sum_\omega \overline{i_{\omega,t}}\mathcal{D}_{\beta_{\mathrm{reg}}}(0\|\sum_k f^{(\mathrm{KL})}_{\omega,k} g_{k,t}) \\ \quad + \mu\|\boldsymbol{F}^{(\mathrm{KL})\mathrm{T}}\boldsymbol{H}\|^2_{\mathrm{Fr}} \qquad (r_t < \tau) \end{cases}, \tag{12}$$

$$s^{(*)}_{\omega,t} = \sum_k f^{(*)}_{\omega,k} g_{k,t} + \sum_n h_{\omega,n} u_{n,t}, \tag{13}$$

$$r_t = \frac{\sum_\omega \overline{i_{\omega,t}}}{\Omega}, \tag{14}$$

where $\boldsymbol{F}^{(\mathrm{KL})}(\in \mathbb{R}^{\Omega\times K}_{\geq 0})$ and $\boldsymbol{F}^{(\mathrm{EUC})}(\in \mathbb{R}^{\Omega\times K}_{\geq 0})$ are the supervised basis matrices trained in advance using KL-divergence and EUC-distance, respectively. Also, $f^{(\mathrm{KL})}_{\omega,k}$ and $f^{(\mathrm{EUC})}_{\omega,k}$ are the entries of $\boldsymbol{F}^{(\mathrm{KL})}$ and $\boldsymbol{F}^{(\mathrm{EUC})}$, respectively, and $* = \{\mathrm{KL}, \mathrm{EUC}\}$. The divergence is determined from r_t and τ in each frame. Therefore, this method can be considered as *multi-divergence-based SNMF* to achieve both optimal separation and extrapolation.

3.3 Auxiliary Function and Update Rules

The update rules based on (11) are obtained by an auxiliary function approach. Similarly to in [12], we can design an upper bound function $\mathcal{J}^+$ using auxiliary variables $\alpha^{(*)}_{k,l,\omega} \geq 0$, $\gamma^{(*)}_{\omega,t,k} \geq 0$, $\delta_{\omega,t,l} \geq 0$, $\varepsilon_1 \geq 0$, $\varepsilon_2 \geq 0$, and $\zeta^{(*)}_{\omega,t} \geq 0$ that satisfy $\sum_\omega \alpha^{(*)}_{k,l,\omega} = 1$, $\sum_k \gamma_{\omega,t,k} = 1$, $\sum_l \delta_{\omega,t,l} = 1$, and $\varepsilon_1 + \varepsilon_2 = 1$ as

$$\mathcal{J} \leq \mathcal{J}^{+} = \sum_{t} \mathcal{J}_t^{+}, \tag{15}$$

$$\mathcal{J}_t \leq \mathcal{J}_t^{+} = \begin{cases} \sum_\omega i_{\omega,t}\left(y_{\omega,t}^2 + v_{\omega,t} + 2w_{\omega,t}\right) \\ \quad +\lambda \sum_\omega \overline{i_{\omega,t}} \mathcal{R}_{\beta_{\mathrm{reg}}}^{(\mathrm{EUC})} \\ \quad +\mu \sum_{k,l,\omega} \left(f_{\omega,k}^{(\mathrm{EUC})2} h_{\omega,l}^2\right) / \alpha_{k,l,\omega}^{(\mathrm{EUC})} & (r_t \geq \tau) \\ \sum_\omega i_{\omega,t}\left(-y_{\omega,t} \sum_{k,l} \gamma_{\omega,t,k}^{(\mathrm{KL})} \delta_{\omega,t,l} Q + C\right) \\ \quad +\lambda \sum_\omega \overline{i_{\omega,t}} \mathcal{R}_{\beta_{\mathrm{reg}}}^{(\mathrm{KL})} \\ \quad +\mu \sum_{k,l,\omega} \left(f_{\omega,k}^{(\mathrm{KL})2} h_{\omega,l}^2\right) / \alpha_{k,l,\omega}^{(\mathrm{KL})} & (r_t < \tau) \end{cases}, \tag{16}$$

where

$$v_{\omega,t} = \sum_k \frac{f_{\omega,k}^{(\mathrm{EUC})2} g_{k,t}^2}{\gamma_{\omega,t,k}^{(\mathrm{EUC})}} + \sum_l \frac{h_{\omega,l} u_{l,t}}{\delta_{\omega,t,l}}, \tag{17}$$

$$\begin{aligned} w_{\omega,t} = {} & \left(\sum_k f_{\omega,k}^{(\mathrm{EUC})} g_{k,t}\right)\left(\sum_l h_{\omega,l} u_{l,t}\right), \\ & - y_{\omega,t} \sum_k f_{\omega,k}^{(\mathrm{EUC})} g_{k,t} - y_{\omega,t} \sum_l h_{\omega,l} u_{l,t}, \end{aligned} \tag{18}$$

$$\mathcal{R}_{\beta_{\mathrm{reg}}}^{(*)} = \begin{cases} \zeta_{\omega,t}^{(*)\beta_{\mathrm{reg}}-1}\left(\sum_k f_{\omega,k}^{(*)} g_{k,t} - \zeta_{\omega,t}^{(*)}\right) + \dfrac{\zeta_{\omega,t}^{(*)\beta_{\mathrm{reg}}}}{\beta_{\mathrm{reg}}} \\ & (\beta_{\mathrm{reg}} < 1) \\ \dfrac{1}{\beta_{\mathrm{reg}}} \sum_k \gamma_{\omega,t,k} \left(\dfrac{f_{\omega,k}^{(*)} g_{k,t}}{\gamma_{\omega,t,k}^{(*)}}\right)^{\beta_{\mathrm{reg}}} & (1 \leq \beta_{\mathrm{reg}}) \end{cases}, \tag{19}$$

$$Q = \varepsilon_1 \log \Phi + \varepsilon_2 \log \Psi, \tag{20}$$

$$\begin{aligned} C = {} & - y_{\omega,t} \sum_{k,l} \gamma_{\omega,t,k}^{(\mathrm{KL})} \delta_{\omega,t,l} \\ & \cdot \left(\log \gamma_{\omega,t,k}^{(\mathrm{KL})} \delta_{\omega,t,l} + \epsilon_1 \log \epsilon_1 + \epsilon_2 \log \epsilon_2\right) \end{aligned} \tag{21}$$

$$\Phi = \delta_{\omega,t,l} f_{\omega,k}^{(\mathrm{KL})} g_{k,t}, \tag{22}$$

$$\Psi = \gamma_{\omega,t,k}^{(\mathrm{KL})} h_{\omega,l} u_{l,t}. \tag{23}$$

The equality in (16) holds if and only if the auxiliary variables are set as follows:

$$\alpha_{k,l,\omega}^{(*)} = \frac{f_{\omega,k}^{(*)} h_{\omega,l}}{\sum_{\omega'} f_{\omega',k}^{(*)} h_{\omega',l}}, \tag{24}$$

$$\gamma_{\omega,t,k}^{(*)} = \frac{f_{\omega,k}^{(*)} g_{k,t}}{\sum_{k'} f_{\omega,k'}^{(*)} g_{k',t}}, \tag{25}$$

$$\delta_{\omega,t,l} = \frac{h_{\omega,l} u_{l,t}}{\sum_{l'} h_{\omega,l'} u_{l',t}}, \tag{26}$$

$$\varepsilon_1 = \frac{\Phi}{\Phi + \Psi}, \tag{27}$$

$$\varepsilon_2 = \frac{\Psi}{\Phi + \Psi}, \tag{28}$$

$$\zeta_{\omega,t}^{(*)} = \sum_k f_{\omega,k}^{(*)} g_{k,t}. \tag{29}$$

The update rules are obtained from the derivative of the upper bound function (15) w.r.t. each objective variable and substitution of the equality conditions as

Table 1 Compositions of musical instruments

Composition	Melody 1	Melody 2	Midrange	Bass
C1	Oboe	Flute	Piano	Trombone
C2	Trumpet	Violin	Harpsichord	Fagotto
C3	Clarinet	Horn	Piano	Cello

Table 2 Spatial conditions of each dataset

Spatial	Measure			
pattern	1st	2nd	3rd	4th
SP1	$\theta=45°$	$\theta=0°$	$\theta=0°$	$\theta=0°$
SP2	$\theta=45°$	$\theta=45°$	$\theta=0°$	$\theta=0°$
SP3	$\theta=45°$	$\theta=45°$	$\theta=45°$	$\theta=0°$
SP4	$\theta=45°$	$\theta=45°$	$\theta=45°$	$\theta=45°$

$$g_{k,t} \leftarrow \begin{cases} \dfrac{g_{k,t} \sum_\omega i_{\omega,t} y_{\omega,t} f_{\omega,k}^{(\mathrm{EUC})}}{\sum_\omega i_{\omega,t} f_{\omega,k}^{(\mathrm{EUC})} s_{\omega,t}^{(\mathrm{EUC})} + \lambda \sum_\omega \overline{i_{\omega,t}} f_{\omega,k}^{(\mathrm{EUC})} \left(\sum_{k'} f_{\omega,k'}^{(\mathrm{EUC})} g_{k',t}\right)^{\beta_{\mathrm{reg}}}} \\ \hfill (r_t \geq \tau) \\ \dfrac{g_{k,t} \sum_\omega i_{\omega,t} y_{\omega,t} f_{\omega,k}^{(\mathrm{KL})} s_{\omega,t}^{(\mathrm{KL})-1}}{\sum_\omega i_{\omega,t} f_{\omega,k}^{(\mathrm{KL})} + \lambda \sum_\omega \overline{i_{\omega,t}} f_{\omega,k}^{(\mathrm{KL})} \left(\sum_{k'} f_{\omega,k'}^{(\mathrm{KL})} g_{k',t}\right)^{\beta_{\mathrm{reg}}}} \\ \hfill (r_t < \tau) \end{cases}, \tag{30}$$

$$h_{\omega,l} \leftarrow \frac{h_{\omega,l} \sum_t i_{\omega,t} y_{\omega,t} u_{l,t} N_{\omega,t}}{\sum_t i_{\omega,t} u_{l,t} D_{\omega,t} + \mu P_{\omega,l}}, \tag{31}$$

$$u_{l,t} \leftarrow \begin{cases} \dfrac{u_{l,t} \sum_\omega i_{\omega,t} y_{\omega,t} h_{\omega,l}}{\sum_\omega i_{\omega,t} h_{\omega,l} s_{\omega,t}^{(\mathrm{EUC})}} & (r_t \geq \tau) \\ \dfrac{u_{l,t} \sum_\omega i_{\omega,t} y_{\omega,t} h_{\omega,l} s_{\omega,t}^{(\mathrm{EUC})-1}}{\sum_\omega i_{\omega,t} h_{\omega,l}} & (r_t < \tau) \end{cases}, \tag{32}$$

where $N_{\omega,t}$, $D_{\omega,t}$, and $P_{\omega,l}$ are given by

$$N_{\omega,t} = \begin{cases} 1 & (r_t \geq \tau) \\ s_{\omega,t}^{(\mathrm{KL})-1} & (r_t < \tau) \end{cases}, \tag{33}$$

$$D_{\omega,t} = \begin{cases} s_{\omega,t}^{(\mathrm{EUC})} & (r_t \geq \tau) \\ 1 & (r_t < \tau) \end{cases}, \tag{34}$$

$$P_{\omega,l} = \begin{cases} \sum_k f_{\omega,k}^{(\mathrm{EUC})} \sum_{\omega'} f_{\omega',k}^{(\mathrm{EUC})} h_{\omega',l} & (r_t \geq \tau) \\ \sum_k f_{\omega,k}^{(\mathrm{KL})} \sum_{\omega'} f_{\omega',k}^{(\mathrm{KL})} h_{\omega',l} & (r_t < \tau) \end{cases}. \tag{35}$$

In total, the update rules of superresolution-based SNMF based on multi-divergence are defined as (30)–(32).

4. Experiments

4.1 Experimental Conditions

To confirm the effectiveness of the proposed algorithm, we compared six methods, namely, simple directional clustering [15], multichannel NMF [11], penalized SNMF (PSNMF) based on KL-divergence and EUC-distance [7], the conventional hybrid method based on KL-divergence and EUC-distance, the online hybrid method described in Sect. 3.2, and the proposed hybrid method that uses multi-divergence. In this experiment, we conducted two experiments to consider artificial signal and real-recorded signal cases. We used stereo signals containing four melody parts (depicted in Fig. 4) with three compositions (C1–C3) of instruments shown in Table 1. These signals were artificially generated by a MIDI synthesizer. In particular, these stereo

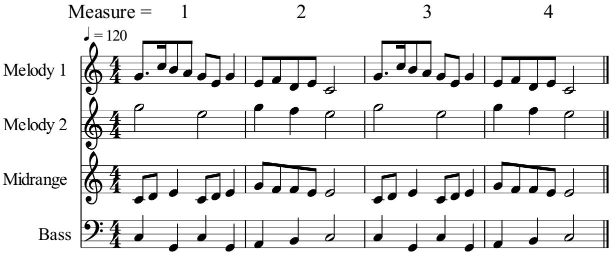

Fig. 4 Scores of each part consisting of four measures.

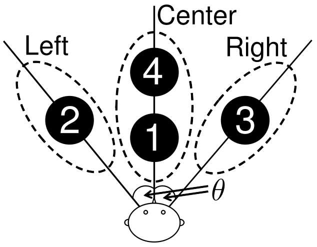

Fig. 5 Panning of four sources, where numbered black circles represent locations of instruments.

signals were mixed down to a monaural format only when we evaluated the separation accuracy of PSNMF because PSNMF is a separation method for a monaural input signal. In the artificial signal case, the observed signals were produced by mixing four sources with the same power. The sources were mixed as shown in Fig. 5, where the target source was always located in the center direction with another interfering source. In addition, we used the same MIDI sounds of the target instruments as supervision for a priori training. The training sounds contained notes over two octaves that covered all the notes of the target signal in the observed signal. The sampling frequency of all signals was 44.1 kHz. The spectrograms were computed using a 92-ms-long rectangular window with a 46-ms overlap shift. The number of iterations used for training and separation was 500, and the threshold value τ was set to 20%. The number of clusters used in directional clustering was 3, the number of a priori bases was 100, and the number of bases for matrix $\boldsymbol{H}$ was 30. The parameters λ and μ were empirically determined.

In the real-recorded signal case, we recorded each instrumental solo signal and the supervision sound, which are the same as those in the artificial signal case, using binaural microphone NEUMANN KU-100 in an experimental room whose reverberation time was 200 ms. The levels of background noise and the sound source measured at the microphone were 37 dB(A) and 60 dB(A). In the center direction, there were two loudspeakers located at a distance of 1.5 m and 2.5 m from the microphone. Also, in each of left- and right-hand sides at $\pm\theta$, the loudspeaker was located 1.5 m apart from the microphone. The observed signals were produced by mixing these signals as the same power. Other conditions were the same as those of the artificial signal case.

We prepared four spatially different dataset patterns of the observed signals, SP1–SP4, as shown in Table 2. In the hybrid method, many chasms were produced by directional clustering in the measures for $\theta = 45°$ compared with those for $\theta = 0°$. Therefore, we can expect that the EUC-distance-based hybrid method is suitable for SP4 rather than for the dataset of SP1.

4.2 Experimental Results

We used the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and sources-to-artifacts ratio (SAR) defined in [18] as the evaluation score. Here, the estimated signal $\hat{s}(t)$ is defined as

$$\hat{s}(t) = s_{\mathrm{target}}(t) + s_{\mathrm{interf}}(t) + s_{\mathrm{artif}}(t), \tag{36}$$

where $s_{\mathrm{target}}(t)$ is the allowable deformation of the target source, $s_{\mathrm{interf}}(t)$ is the allowable deformation of the sources that account for the interferences of the undesired sources, and $s_{\mathrm{artif}}(t)$ is an *artifact* term that may correspond to the artifacts of the separation algorithm, such as musical noise, or simply undesirable deformation induced by the nonlinear property of the separation algorithm. The formulae for SDR, SIR, and SAR are defined as

$$\mathrm{SDR} = 10\log_{10}\frac{\sum_t s_{\mathrm{target}}(t)^2}{\sum_t \{e_{\mathrm{interf}}(t) + e_{\mathrm{artif}}(t)\}^2}, \tag{37}$$

$$\mathrm{SIR} = 10\log_{10}\frac{\sum_t s_{\mathrm{target}}(t)^2}{\sum_t e_{\mathrm{interf}}(t)^2}, \tag{38}$$

$$\mathrm{SAR} = 10\log_{10}\frac{\sum_t \left\{s_{\mathrm{target}}(t) + e_{\mathrm{interf}}(t)\right\}^2}{\sum_t e_{\mathrm{artif}}(t)^2}. \tag{39}$$

SDR indicates the quality of the separated target sound, which includes the degree of separation (SIR) and the absence of artificial distortion (SAR).

Figures 6 and 7 show the average scores for each method and each dataset pattern, where four instruments are shuffled with 12 combinations of each of the compositions C1–C3. Therefore, these results are the averages of 36 input signals. In addition, the scores of PSNMF are the same for any datasets because the input signals for PSNMF are mixed down to a monaural format. From these results, the methods except the hybrid method cannot obtain enough separation performance. The KL-divergence-based hybrid method achieves high separation accuracy for the dataset of spatial patterns SP1 and SP2 because these signals do not have many spectral chasms. On the other hand, the EUC-divergence-based hybrid method achieves high separation accuracy for SP4. This dataset has many spectral chasms because the signals are always mixed with a wide panning angle ($\theta = 45°$), which yields many chasms, and high extrapolation ability is required. The proposed hybrid method with multi-divergence can always achieve better separation for any dataset regardless of whether or not many chasms exist. This is because the proposed method selects the appropriate divergence and can automatically apply the optimal divergence to each time frame.

5. Conclusion

We propose a new multi-divergence method to separate the target signal using the optimal divergence. The proposed method adapts the divergence in each frame to the optimal one using a threshold value for the rate of chasms to separate and extrapolate the target signal with high accuracy. Experimental results show that our proposed method can always achieve high separation accuracy under all spatial conditions.

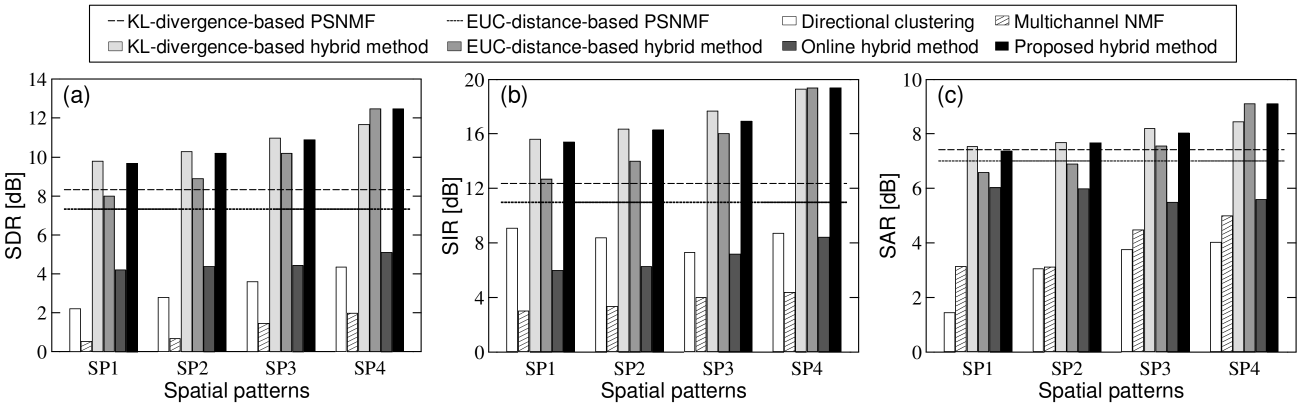

Fig. 6 Average scores of each method and each spatial condition in artificial signal case: (a) shows SDR, (b) shows SIR, and (c) shows SAR.

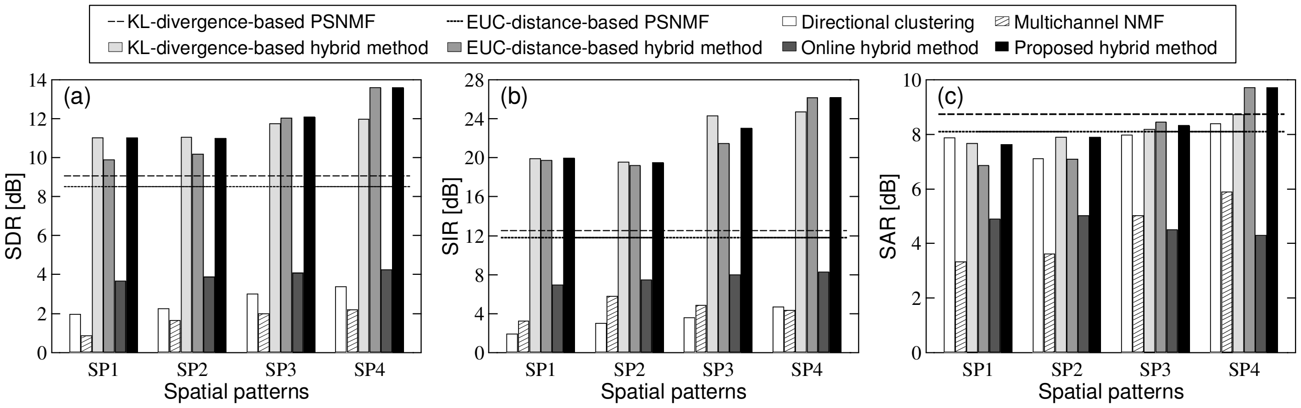

Fig. 7 Average scores of each method and each spatial condition in real-recorded signal case: (a) shows SDR, (b) shows SIR, and (c) shows SAR.

References

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

[2] 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.

[3] S. A. Raczynski, N. Ono, S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," *Proc. 8th International Conference on Music Information Retrieval*, pp.381–386, 2007.

[4] W. Wang, A. Cichocki, J. A. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance," *Signal Processing*, vol.57, no.7, pp.2858–2864, 2009.

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

[6] P. Smaragdis, B. Raj, M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. 7th International Conference on Independent Component Analysis and Signal Separation*, pp.414–421, 2007.

[7] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, K. Kondo, "Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing," *Proc. IEEE International Symposium on Signal Processing and Information Technology*, pp.392–397, 2013.

[8] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, 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, 2014 (in press).

[9] D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, Y. Takahashi, "Music signal separation by supervised nonnegative matrix factorization with basis deformation," *Proc. IEEE 18th International Conference on Digital Signal Processing*, T3P(C)-1, 2013.

[10] A. Ozerov, C. Fevotte, "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.

[11] H. Sawada, H. Kameoka, S. Araki, N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.261–264, 2012.

[12] D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, Y. Takahashi, "Superresolution-based stereo signal separation via supervised nonnegative matrix factorization," *Proc. IEEE 18th International Conference on Digital Signal Processing*, T3C-2, 2013.

[13] D. FitzGerald, M. Cranitch, E. Coyle, "On the use of the beta divergence for musical source separation," *Proc. Irish Signals and Systems Conference*, 2009.

[14] S. Miyabe, K. Masatoki, H. Saruwatari, K. Shikano, T. Nomura, "Temporal quantization of spatial information using directional clustering for multichannel audio coding," *Proc. WASPAA*, pp.261–264, 2009.

[15] S. Araki, H. Sawada, R. Mukai, S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol.87, no.8, pp.1833–1847, 2007.

[16] D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, Y. Takahashi, "Superresolution-based stereo signal separation with regularization of supervised basis extrapolation," *Proc. 3DSA 2013*, no.S10-4, 2013.

[17] S. Eguchi, Y. Kano, "Robustifying maximum likelihood estimation," *Technical Report of Institute of Statistical Mathematics*, 2001.

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