Hybrid Multichannel Signal Separation Using Supervised Nonnegative Matrix Factorization with Spectrogram Restoration

Daichi Kitamura*, Hiroshi Saruwatari†, Satoshi Nakamura‡, Yu Takahashi§,
Kazunobu Kondo§ and Hirokazu Kameoka†
*The Graduate University for Advanced Studies, 2–1–2 Hitotsubashi, Chiyoda, Tokyo, 101–8430, Japan.
†The University of Tokyo, 7–3–1 Hongo, Bunkyo, Tokyo, 113–8656, Japan.
‡Nara Institute of Science and Technology, 8916–5 Takayama-cho, Ikoma, Nara, 630–0192, Japan.
§Yamaha Corporation, 203 Matsunokijima, Iwata, Shizuoka, 438–0192, Japan.

***Abstract*—In this paper, we propose a new hybrid method that concatenates directional clustering and advanced nonnegative matrix factorization (NMF) for the purpose of the specific sound extraction from the multichannel music signal. Multichannel music signal separation technology is aimed to extract a specific target signal from observed multichannel signals that contain multiple instrumental sounds. In the previous studies, various methods using NMF have been proposed, but they remain many problems, e.g., poor convergence in update rules in NMF and lack of robustness. To solve these problems, we propose a new supervised NMF (SNMF) with spectrogram restoration and its hybrid method that concatenates the proposed SNMF after directional clustering. Via extrapolation of supervised spectral bases, the proposed SNMF attempts both target signal separation and reconstruction of the lost target components, which are generated by preceding directional clustering. In addition, we theoretically reveal the trade-off between separation and extrapolation abilities and propose a new scheme for multi-divergence, where optimal divergence can be automatically changed in each time frame according to the local spatial conditions. The results of an evaluation experiment show that our proposed hybrid method outperforms the conventional music signal separation methods.**

I. Introduction

Music signal separation technologies have attracted considerable interest and been intensively studied [2], [3] in recent years. These techniques are underdetermined separation problems because almost all musical tunes are provided in a stereo format and the number of sources is greater than two. As a means of addressing underdetermined signal separation, in recent years, nonnegative matrix factorization (NMF) [4], which is a type of sparse representation algorithm, has received much attention. NMF for acoustical signals decomposes an input spectrogram into the product of a spectral basis matrix and its activation matrix. The methods of signal separation based on NMF are roughly classified into unsupervised and supervised algorithms. The former method attempts separation without using any training sequences, instead being subjected to various constraints, as proposed in [5], [6]. However, these techniques have difficulty in clustering the decomposed spectral bases into a specific target sound because the entire procedure should be carried out in a blind fashion. To solve this problem, supervised NMF (SNMF) has been proposed [7]. This method includes a priori training, which requires some sound samples of a target instrument, and separate the target signal using supervised bases. SNMF can extract the target signal to some extent, particularly in the case of a small number of sources. However, for a mixture consisting of many sources, the extraction performance is markedly degraded because of the existence of instruments with similar timbre.

To apply NMF-based separation methods to multichannel signals, multichannel NMF has been proposed as an unsupervised separation method [8], [9]. This method is a natural extension of NMF for a stereo or multichannel signal and is a unified method that addresses the spatial and spectral separation problems simultaneously. However, such unsupervised separation is a difficult problem, even if the signal has multichannel components, because the decomposition is underspecified. Hence, these algorithms involve strong dependence on initial values and lack robustness. For multichannel signal separation, directional clustering has also been proposed as an unsupervised method [10], [11]. This method quantizes directional information via time-frequency binary masking. However, there is an inherent problem that sources located in the same direction cannot be separated using only the directional information.

To cope with these problems, in this paper, we propose a new SNMF with spectrogram restoration and its hybrid method that concatenates the proposed SNMF after directional clustering. Via extrapolation of supervised spectral bases, this SNMF with spectrogram restoration attempts both target signal separation and reconstruction of the lost target components, which are generated by preceding binary masking performed in directional clustering.

Next, we provide a theoretical analysis of basis extrapolation ability and reveal the mechanism of marked shift of optimal divergence in SNMF with spectrogram restoration and trade-off between separation and extrapolation abilities. Evaluation experiment of the separation using artificial and

The contents in this paper have been submitted to another journal [1].

978-616-361-823-8

real-recorded music signals show the effectiveness of the proposed hybrid method.

Finally, on the basis of the above-mentioned findings, we propose a new scheme for frame-wise divergence selection in the proposed hybrid method to separate the target signal using optimal multi-divergence. The results of an evaluation experiment show that the proposed hybrid method with multi-divergence can always achieve high performance under any spatial conditions, indicating the improvement in robustness of the proposed method.

II. Conventional Signal Separation Methods

A. Conventional Single-Channel Signal Separation Methods

1) Overview of NMF: NMF is a type of sparse representation algorithm that decomposes a nonnegative matrix into two nonnegative matrices as

$$\boldsymbol{X} \simeq \boldsymbol{V}\boldsymbol{W}, \tag{1}$$

where $\boldsymbol{X}(\in \mathbb{R}_{\geq 0}^{M\times N})$ is an observed nonnegative matrix, which is an amplitude spectrogram for applying NMF to the acoustic signal; $\boldsymbol{V}(\in \mathbb{R}_{\geq 0}^{M\times D})$ is often called the *basis matrix*, which includes bases (frequently-appearing spectral patterns in $\boldsymbol{X}$) as column vectors; and $\boldsymbol{W}(\in \mathbb{R}_{\geq 0}^{D\times N})$ is often called the *activation matrix*, which involves activation information of each basis of $\boldsymbol{V}$. In addition, M and N are the numbers of rows and columns of $\boldsymbol{X}$, and D is the number of bases of $\boldsymbol{V}$. Figure 1 depicts the decomposition model of NMF, where the number of bases D equals two. In this figure, the basis matrix includes two types of spectral patterns as the bases to represent the observed matrix using time varying gains in the activation matrix. In the decomposition of NMF, a cost function is defined to optimize the variables $\boldsymbol{V}$ and $\boldsymbol{W}$ using an arbitrary divergence between $\boldsymbol{X}$ and $\boldsymbol{V}\boldsymbol{W}$. The following equation represents the cost function of NMF:

$$\mathcal{J}_{\mathrm{NMF}} = \mathcal{D}(\boldsymbol{X}\|\boldsymbol{V}\boldsymbol{W}), \tag{2}$$

where $\mathcal{D}(\cdot\|\cdot)$ is an arbitrary distance function, e.g., Itakura-Saito divergence (*IS-divergence*), generalized Kullback-Leibler divergence (*KL-divergence*), and Euclidean distance (*EUC-distance*). In this study, we use the following generalized divergence called β-divergence [12] in the cost function:

$$\mathcal{D}_\beta(\boldsymbol{B}\|\boldsymbol{A}) = \begin{cases} \sum_{i,j}\left\{\dfrac{b_{ij}^\beta}{\beta(\beta-1)}+\dfrac{a_{ij}^\beta}{\beta}-\dfrac{b_{ij}a_{ij}^{\beta-1}}{\beta-1}\right\} & (\beta\in\mathbb{R}_{\setminus\{0,1\}}) \\ \sum_{i,j}\left\{b_{ij}\log\dfrac{b_{ij}}{a_{ij}}+a_{ij}-b_{ij}\right\} & (\beta=1) \\ \sum_{i,j}\left\{\dfrac{b_{ij}}{a_{ij}}-\log\dfrac{b_{ij}}{a_{ij}}-1\right\} & (\beta=0) \end{cases}, \tag{3}$$

where $\boldsymbol{A}(\in\mathbb{R}^{I\times J})$ and $\boldsymbol{B}(\in\mathbb{R}^{I\times J})$ are matrices whose entries are a_{ij} and b_{ij}, respectively. This divergence is a family of cost functions parameterized by a single shape parameter β that takes IS-divergence, KL-divergence, and EUC-distance as special cases ($\beta=0, 1$, and 2, respectively).

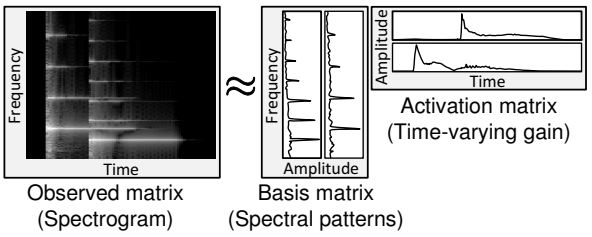

Fig. 1. Decomposition model of simple NMF.

The multiplicative update rules for $\boldsymbol{V}$ and $\boldsymbol{W}$ that minimize the cost function based on β-divergence are given by [13]

$$v_{md} \leftarrow v_{md}\left(\frac{\sum_n x_{mn}w_{dn}\left(\sum_d v_{md}w_{dn}\right)^{\beta-2}}{\sum_n w_{dn}\left(\sum_d v_{md}w_{dn}\right)^{\beta-1}}\right)^{\varphi(\beta)}, \tag{4}$$

$$w_{dn} \leftarrow w_{dn}\left(\frac{\sum_m v_{md}x_{mn}\left(\sum_d v_{md}w_{dn}\right)^{\beta-2}}{\sum_m v_{md}\left(\sum_d v_{md}w_{dn}\right)^{\beta-1}}\right)^{\varphi(\beta)}, \tag{5}$$

where x_{mn}, v_{md}, and w_{dn} are the nonnegative entries of matrices $\boldsymbol{X}$, $\boldsymbol{V}$, and $\boldsymbol{W}$, respectively. In addition, $\varphi(\beta)$ is given by

$$\varphi(\beta) = \begin{cases} (2-\beta)^{-1} & (\beta<1) \\ 1 & (1\leq\beta\leq 2) \\ (\beta-1)^{-1} & (\beta>2) \end{cases}. \tag{6}$$

We can optimize $\boldsymbol{V}$ and $\boldsymbol{W}$ by some iterations of these update rules. The convergence of these update rules is theoretically proven for any real-valued β.

2) SNMF: The signal separation using NMF is achieved by extracting only the target spectral bases. However, such unsupervised approaches have difficultly in clustering the decomposed spectral patterns into a specific target instruments. Furthermore, each basis may be forced to include a multi-instrumental spectral pattern. To solve this problem, SNMF has been proposed [7]. These supervised scheme consists of two processes, namely, a priori training and observed signal separation.

In SNMF, as the supervision, a priori spectral patterns (bases) should be trained in advance to achieve signal separation. Hereafter, we assume that we can obtain specific solo-played instrumental sounds, which is the target of the separation task. The trained bases are constructed by NMF as

$$\boldsymbol{Y}_{\mathrm{target}} \simeq \boldsymbol{F}\boldsymbol{Q}, \tag{7}$$

where $\boldsymbol{Y}_{\mathrm{target}}(\in\mathbb{R}_{\geq 0}^{\Omega\times T_s})$ is an amplitude spectrogram of the specific instrumental signal for training, $\boldsymbol{F}(\in\mathbb{R}_{\geq 0}^{\Omega\times K})$ is a nonnegative matrix that involves bases of the target signal as column vectors, and $\boldsymbol{Q}(\in\mathbb{R}_{\geq 0}^{K\times T_s})$ is a nonnegative matrix that corresponds to the activation of each basis of $\boldsymbol{F}$. In addition, Ω is the number of frequency bins, T_s is the number of frames of the training signal, and K is the number of bases. Therefore, the basis matrix $\boldsymbol{F}$ constructed by (7) is the supervision of the target instrumental spectra.

The following equation represents the decomposition model in separation process with trained supervision $\boldsymbol{F}$:

$$\boldsymbol{Y} \simeq \boldsymbol{FG} + \boldsymbol{HU}, \tag{8}$$

where $\boldsymbol{Y}(\in \mathbb{R}_{\geq 0}^{\Omega \times T})$ is an observed spectrogram, $\boldsymbol{G}(\in \mathbb{R}_{\geq 0}^{K \times T})$ is an activation matrix that corresponds to $\boldsymbol{F}$, $\boldsymbol{H}(\in \mathbb{R}_{\geq 0}^{\Omega \times L})$ is the residual spectral patterns that cannot be expressed by $\boldsymbol{FG}$, and $\boldsymbol{U}(\in \mathbb{R}_{\geq 0}^{L \times T})$ is an activation matrix that corresponds to $\boldsymbol{H}$. Moreover, T is the number of frames of the observed signal 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. Hence, ideally, $\boldsymbol{FG}$ represents the target instrumental components, and $\boldsymbol{HU}$ represents other different components from the target sounds after the decomposition. This supervised method can separate the target signal to some extent, particularly in the case of a small number of sources. However, for the case of a mixture consisting of many sources, such as more realistic musical tunes, the source extraction performance is markedly degraded because of the existence of instruments with similar timbre.

B. Conventional Multichannel Signal Separation Methods

1) Directional Clustering: Decomposition methods employing directional information for the multichannel signal have also been proposed as unsupervised separation techniques [10], [11]. These methods quantize directional information via time-frequency binary masking under the assumption that the sources are completely sparse (double disjoint) in the time-frequency domain. Such directional clustering works well, even in an underdetermined situation. However, there is an inherent problem that sources located in the same direction cannot be separated using the directional information. Furthermore, the extracted signal is likely to be distorted because of the effect of binary masking.

2) Multichannel NMF: Multichannel NMF, which is a natural extension of NMF for a stereo or multichannel music signal, has been proposed as an unsupervised signal separation method [8], [9]. These algorithms employ Hermitian positive definite matrix that models the spatial property of each NMF basis and each frequency bin. Therefore, multichannel NMF utilizes a frequency-wise transfer function between signal source and microphone as a cue for basis clustering. However, such unsupervised separation is a difficult problem, even if the signal has multichannel components, because the decomposition is underspecified. Hence, these algorithms involve strong dependence on initial values and lack robustness.

III. SNMF with Spectrogram Restoration and Its Hybrid Method

A. SNMF with Spectrogram Restoration

1) Motivation and Strategy: To separate the target source utilizing directional information, we can guess a hybrid method that concatenates SNMF after directional clustering (hereafter referred to as *naive hybrid method*). This hybrid method can effectively extract the target instrument because the directionally clustered signal contains only few instruments. Moreover, the residual interfering signal in the same direction can be removed by SNMF.

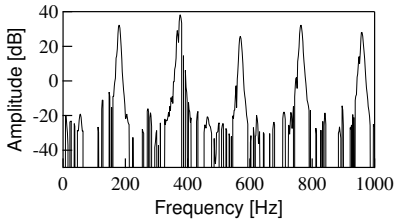

Fig. 2. Example of spectrum of signal separated by directional clustering.

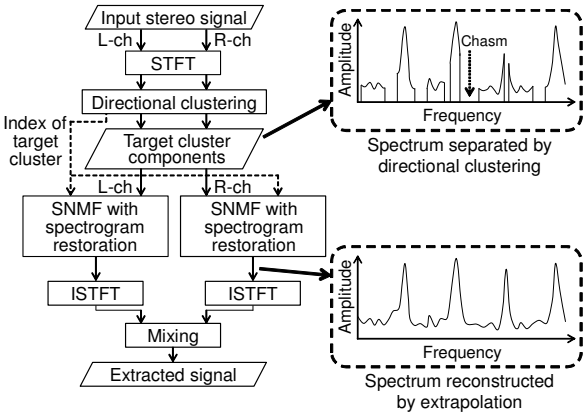

Fig. 3. Signal flow of proposed hybrid method; SNMF with spectrogram restoration concatenates after directional clustering.

However, such naive hybrid method has a problem that the extracted signal may suffer from the generation of considerable distortion. This is due to the binary masking in directional clustering. The signal in the target direction, which is obtained by directional clustering, has many spectral chasms because the assumption of sparseness in the time-frequency domain does not always hold completely. In other words, the resolution of the spectrogram clustered as the target-direction component is degraded by time-frequency binary masking. Figure 2 shows an example of the spectrum of a signal separated by directional clustering. The obtained spectrum has many chasms owing to the binary masking. These spectral losses may deteriorate the performance of separation because SNMF is forced to incorrectly fit these spectral chasms using supervised bases. To solve this problem, in this section, we propose a new SNMF with spectrogram restoration as an alternative to the conventional SNMF for the hybrid method.

Figure 3 shows a signal flow of the proposed hybrid method that includes SNMF with spectrogram restoration. The algorithm of SNMF with spectrogram restoration utilizes index information determined in directional clustering. For example, if the target instrument is localized in the center cluster along with the interference, SNMF is only applied to the existing center components using index information (active binary mask). Therefore, the spectrogram of the target instrument is reconstructed using more matched bases because spectral chasms are treated as *unseen*, and these chasms have no impact on the cost function in SNMF with spectrogram restoration.

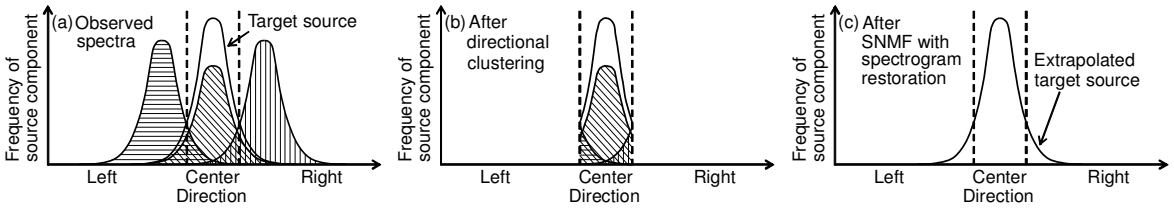

Fig. 4. Directional source distribution of (a) observed stereo signal, (b) separated components in center cluster, and (c) component separated and extrapolated by spectrogram restoration.

In addition, the components of the target instrument lost after directional clustering can be extrapolated using the supervised bases. In other words, the deteriorated target spectrogram is recovered with the spectrogram restoration by the supervised basis extrapolation.

To illustrate the separation mechanism step by step, Fig. 4 (a) shows the configuration of source components in the stereo signal, (b) shows the separated components that are clustered around the center direction by directional clustering, and (c) shows the separated target component obtained by SNMF with spectrogram restoration. In Fig. 4 (a), the source components are distributed in all directions with some overlapping. After directional clustering (Fig. 4 (b)), the center sources lose some of their components (i.e., the tails on both sides), and the other source components leak in the center cluster. After SNMF with spectrogram restoration, the proposed algorithm restores the lost components using the supervised bases (Fig. 4 (c)).

However, this basis extrapolation includes an underlying problem. If the time-frequency spectra are almost unseen in the spectrogram, which means that the indexes are almost zero, a large extrapolation error may occur. Then, incorrect bases are chosen and fitted to a small number of spectral grids by incorrectly modifying the activation matrix $\boldsymbol{G}$. In the worst case, the activation matrix $\boldsymbol{G}$ contains very large values and the extracted signal is overloaded. To avoid this, we should add a new penalty term in the cost function, as described in the next section.

2) Cost Function and Update Rules: In this section, we derive the update rules of SNMF with spectrogram restoration based on β-divergence. Here, the index matrix $\boldsymbol{I}(\in \mathbb{R}^{\Omega\times T}_{\{0,1\}})$ is obtained from the binary masking preceding the directional clustering. This index matrix has specific entries of unity or zero, which indicates whether or not each grid of the spectrogram belongs to the target directional cluster. The cost function in SNMF with spectrogram restoration is defined using the index matrix $\boldsymbol{I}$ as

$$\begin{aligned}\mathcal{J}(\Theta)=&\textstyle\sum_{\omega,t} i_{\omega t}\mathcal{D}_{\beta}(y_{\omega t}\|\sum_k f_{\omega k}g_{kt}+\sum_l h_{\omega l}u_{lt})\\ &+\lambda\textstyle\sum_{\omega,t}\overline{i_{\omega t}}\mathcal{D}_{\beta_{\mathrm{R}}}(0\|\sum_k f_{\omega k}g_{kt})+\mu\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|^2_{\mathrm{F}},\end{aligned}\quad (9)$$

where $y_{\omega t}$, $f_{\omega k}$, g_{kt}, $h_{\omega l}$, and u_{lt} are the nonnegative entries of the matrices $\boldsymbol{Y}$, $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$, respectively, $\Theta = \{\boldsymbol{G},\boldsymbol{H},\boldsymbol{U}\}$ is the set of objective variables, $i_{\omega t}$ is an entry of the index matrix $\boldsymbol{I}$, λ and μ are the weighting parameters for each term, and $\overline{i_{\omega t}}$ represents the binary complement of each entry in the index matrix. The first term represents the main cost of separation in SNMF. Since the divergence $\mathcal{D}_{\beta}(\cdot\|\cdot)$ is only defined in grids whose index is one, the chasms in the spectrogram are ignored in this SNMF decomposition. The second term forces the minimization of the value of $\sum_k f_{\omega k}g_{kt}$. Hence, the supervised bases are chosen so as to minimize the scale of $\boldsymbol{FG}$ in proportion to the number of zeros in the index matrix $\boldsymbol{I}$ in each frame to avoid the extrapolation error. In other words, this penalty term regulates the extrapolation. In addition, the third penalty term forces the other bases $\boldsymbol{H}$ to become as different as possible from the supervised bases $\boldsymbol{F}$ and can improve its separation performance [14].

The update rules based on (9) are obtained by the auxiliary function approach, similarly to [13]. Here, we can rewrite the cost function (9) using β-divergence as

$$\mathcal{J}(\Theta) = \mathcal{J}_1 + \lambda\mathcal{J}_2 + \mu\mathcal{J}_3, \quad (10)$$

$$\mathcal{J}_1 = \textstyle\sum_{\omega,t} i_{\omega t}\left(z^{\beta}_{\omega t}/\beta - y_{\omega t}z^{\beta-1}_{\omega t}/(\beta-1)\right), \quad (11)$$

$$\mathcal{J}_2 = \textstyle\sum_{\omega,t}\overline{i_{\omega t}}\left(\sum_k f_{\omega k}g_{kt}\right)^{\beta_{\mathrm{R}}}/\beta_{\mathrm{R}}, \quad (12)$$

$$\mathcal{J}_3 = \textstyle\sum_{k,l}\left(\sum_{\omega} f_{\omega k}h_{\omega l}\right)^2, \quad (13)$$

where constant terms are omitted and

$$z_{\omega t} = \textstyle\sum_k f_{\omega k}g_{kt} + \sum_l h_{\omega l}u_{lt}. \quad (14)$$

First, we define the upper bound function for $\mathcal{J}_1$. The first term of $\mathcal{J}_1$ is convex for $\beta \geq 1$ and concave for $\beta < 1$, and the second term is convex for $\beta \geq 2$ and concave for $\beta < 2$. Applying Jensen's inequality to the convex function and the tangent line inequality to the concave function, we can define the upper bound function $\mathcal{J}_1^+$ using auxiliary variables $\alpha_{\omega tk} \geq 0$, $\gamma_{\omega tl} \geq 0$, $\eta_1 \geq 0$, $\eta_2 \geq 0$, and $\sigma_{\omega t}$ that satisfy $\sum_k \alpha_{\omega tk} = 1$, $\sum_l \gamma_{\omega tl} = 1$, and $\eta_1 + \eta_2 = 1$ as

$$\mathcal{J}_1 \leq \mathcal{J}_1^+ = \textstyle\sum_{\omega,t} i_{\omega t}\mathcal{P}^{(\beta)}_{\omega t}, \quad (15)$$

where

$$\mathcal{P}^{(\beta)}_{\omega t} = \begin{cases}\mathcal{N}^{(\beta)}_{\omega t} - y_{\omega t}\mathcal{M}^{(\beta-1)}_{\omega t} & (\beta<1)\\ \mathcal{M}^{(\beta)}_{\omega t} - y_{\omega t}\mathcal{M}^{(\beta-1)}_{\omega t} & (1\leq\beta\leq 2)\\ \mathcal{M}^{(\beta)}_{\omega t} - y_{\omega t}\mathcal{N}^{(\beta-1)}_{\omega t} & (\beta>2)\end{cases}, \quad (16)$$

$$\mathcal{M}^{(\beta)}_{\omega t} = \frac{1}{\beta}\textstyle\sum_k \frac{\left(f_{\omega k}g_{kt}\right)^{\beta}}{\left(\alpha_{\omega tk}\eta_1\right)^{\beta-1}} + \frac{1}{\beta}\sum_l \frac{\left(h_{\omega l}u_{lt}\right)^{\beta}}{\left(\gamma_{\omega tl}\eta_2\right)^{\beta-1}}, \quad (17)$$

$$\mathcal{N}^{(\beta)}_{\omega t} = \sigma^{\beta-1}_{\omega t}\left(z_{\omega t} - \sigma_{\omega t}\right) + \sigma^{\beta}_{\omega t}/\beta. \quad (18)$$

Second, we define the upper bound function for $\mathcal{J}_2$. This term is convex for $\beta_{\mathrm{R}} \geq 1$ and concave for $\beta_{\mathrm{R}} < 1$. Similarly to (15)–(18), we can define the upper bound function $\mathcal{J}_2^+$ using auxiliary variables $\alpha_{\omega tk}$ and $\rho_{\omega t}$ as

$$\mathcal{J}_2 \leq \mathcal{J}_2^+ = \textstyle\sum_{\omega,t}\overline{i_{\omega t}}\mathcal{S}^{(\beta_{\mathrm{R}})}_{\omega t}, \quad (19)$$

where

$$\mathcal{S}_{\omega t}^{(\beta_{\mathrm{R}})}=\begin{cases}\rho_{\omega t}^{\beta_{\mathrm{R}}-1}\left(\sum_k f_{\omega k}g_{kt}-\rho_{\omega t}\right)+\dfrac{\rho_{\omega t}^{\beta_{\mathrm{R}}}}{\beta_{\mathrm{R}}} & (\beta_{\mathrm{R}}<1)\\ \dfrac{1}{\beta_{\mathrm{R}}}\sum_k \alpha_{\omega tk}\left(\dfrac{f_{\omega k}g_{kt}}{\alpha_{\omega tk}}\right)^{\beta_{\mathrm{R}}} & (1\leq\beta_{\mathrm{R}})\end{cases}. \quad (20)$$

Third, we define the upper bound function for $\mathcal{J}_3$ using auxiliary variables $\delta_{kl\omega}\geq 0$ that satisfy $\sum_\omega \delta_{kl\omega}=1$ as

$$\mathcal{J}_3\leq\mathcal{J}_3^{+}=\sum_{k,l,\omega}\frac{f_{\omega k}^2 h_{\omega l}^2}{\delta_{kl\omega}}. \quad (21)$$

Finally, using (15), (19), and (21), we can define the upper bound function $\mathcal{J}^{+}(\Theta,\hat{\Theta})$ as

$$\mathcal{J}(\Theta)\leq\mathcal{J}^{+}(\Theta,\hat{\Theta})=\mathcal{J}_1^{+}+\lambda\mathcal{J}_2^{+}+\mu\mathcal{J}_3^{+}, \quad (22)$$

where $\hat{\Theta}$ is the set of auxiliary variables. The update rules with respect to each variable are determined by setting the gradient to zero.

From $\partial\mathcal{J}^{+}(\Theta,\hat{\Theta})/\partial g_{kt}=0$, we obtain

$$\sum_\omega i_{\omega t}\left(\mathcal{V}_\beta-\mathcal{W}_\beta\right)+\lambda\mathcal{X}_{\beta_{\mathrm{R}}}=0, \quad (23)$$

where

$$\mathcal{V}_\beta=\begin{cases}\sigma_{\omega t}^{\beta-1}f_{\omega k} & (\beta<1)\\ g_{kt}^{\beta-1}\left(\alpha_{k\omega t}\eta_1\right)^{1-\beta}f_{\omega k}^{\beta} & (1\leq\beta)\end{cases}, \quad (24)$$

$$\mathcal{W}_\beta=\begin{cases}y_{\omega t}g_{kt}^{\beta-2}\left(\alpha_{k\omega t}\eta_1\right)^{2-\beta}f_{\omega k}^{\beta-1} & (\beta\leq 2)\\ y_{\omega t}\sigma_{\omega t}^{\beta-2}f_{\omega k} & (2<\beta)\end{cases}, \quad (25)$$

$$\mathcal{X}_{\beta_{\mathrm{R}}}=\begin{cases}\sum_\omega \overline{i_{\omega t}}\rho_{\omega t}^{\beta_{\mathrm{R}}-1}f_{\omega k} & (\beta_{\mathrm{R}}<1)\\ \sum_\omega \overline{i_{\omega t}}f_{\omega k}\left(\dfrac{f_{\omega k}g_{kt}}{\alpha_{\omega tk}}\right)^{\beta_{\mathrm{R}}-1} & (1\leq\beta_{\mathrm{R}})\end{cases}. \quad (26)$$

By solving (23) for g_{kt} under the nonnegativity and substituting the equality condition for each auxiliary variable, we can obtain the update rule of g_{kt} as follows:

$$g_{kt}\leftarrow g_{kt}\left(\frac{\sum_\omega i_{\omega t}y_{\omega t}f_{\omega k}z_{\omega t}^{\beta-2}}{\sum_\omega i_{\omega t}f_{\omega k}z_{\omega t}^{\beta-1}+\lambda R_G}\right)^{\varphi(\beta)}, \quad (27)$$

where R_G is given by

$$R_G=\sum_\omega \overline{i_{\omega t}}f_{\omega k}\left(\sum_{k'}f_{\omega k'}g_{k't}\right)^{\beta_{\mathrm{R}}-1}. \quad (28)$$

The update rules of the other variables are similarly obtained as follows:

$$h_{\omega l}\leftarrow h_{\omega l}\left(\frac{\sum_t i_{\omega t}y_{\omega t}u_{lt}z_{\omega t}^{\beta-2}}{\sum_t i_{\omega t}u_{lt}z_{\omega t}^{\beta-1}+2\mu R_H}\right)^{\varphi(\beta)}, \quad (29)$$

$$u_{lt}\leftarrow u_{lt}\left(\frac{\sum_\omega i_{\omega t}y_{\omega t}h_{\omega l}z_{\omega t}^{\beta-2}}{\sum_\omega i_{\omega t}h_{\omega l}z_{\omega t}^{\beta-1}}\right)^{\varphi(\beta)}, \quad (30)$$

where R_H is given by

$$R_H=\sum_k f_{\omega k}\sum_{\omega'}f_{\omega' k}h_{\omega' l}. \quad (31)$$

The convergence of these update rules is theoretically proven for any real-valued β and β_{R}.

B. Theoretical Analysis of Basis Extrapolation Based on Generation Model

1) Optimal Divergence for Basis Extrapolation and Generation Model: The proposed method attempts both signal *separation* and basis *extrapolation* using the supervised bases $\boldsymbol{F}$. In previous studies, the analysis of optimal divergence only for signal separation has been discussed [14], [15]. However, there has been no discussion on the optimal divergence for the extrapolation techniques using NMF. In this section, we analyze the extrapolation ability based on a statistical generation model of the observed data $\boldsymbol{Y}$, and determine the optimal divergence for basis extrapolation w.r.t. various β and β_{R} values.

In NMF decomposition, the minimization of β-divergence between $\boldsymbol{Y}$ and $\boldsymbol{FG}$ corresponds to a log-likelihood maximization under the assumption of the generation model of $\boldsymbol{Y}$ for each β [16]. The minimization of $\mathcal{D}_\beta(y_{\omega t}\|\vartheta)$ is equivalent to the maximization of $\exp(-\mathcal{D}_\beta(y_{\omega t}\|\vartheta))$, where $\vartheta=\sum_k f_{\omega k}g_{kt}$ represents a parameter of the maximum likelihood estimation. A probability density function (p.d.f.) is given by

$$y_{\omega t}\sim p\left(y_{\omega t}\right)=\begin{cases}\dfrac{1}{\vartheta_1}\exp\left(-\dfrac{y_{\omega t}}{\vartheta_1}\right) & (\beta=0)\\ \dfrac{{\vartheta_2}^{y_{\omega t}}}{\Gamma\left(y_{\omega t}+1\right)}\exp\left(-\vartheta_2\right) & (\beta=1)\\ \dfrac{1}{\sqrt{2\pi\vartheta_3}}\exp\left(-\dfrac{\left(y_{\omega t}-\vartheta_4\right)^2}{2{\vartheta_3}^2}\right) & (\beta=2)\\ C\exp\left(\dfrac{{\vartheta_5}^{\beta-1}y_{\omega t}}{\beta-1}\right) & (\beta\geq 3)\end{cases}, \quad (32)$$

where $\Gamma(\cdot)$ is a gamma function. These generation models of $\beta=0$, 1, and 2 are equivalent to exponential distribution, Poisson distribution, and Gaussian distribution, respectively. The generation models for $\beta\geq 3$ correspond to a distribution in which the probability increases exponentially with increasing $y_{\omega t}$. Strictly, this distribution is not a p.d.f. because it diverges when $y_{\omega t}$ increases. Thus, we set the upper bound of $y_{\omega t}$ to a constant M and define the corresponding p.d.f. with normalization coefficient C, which is given by

$$C={\vartheta_5}^{\beta-1}\left(\beta-1\right)^{-1}\left(\exp\left(\frac{{\vartheta_5}^{\beta-1}}{\beta-1}M\right)-1\right)^{-1}. \quad (33)$$

Using (32), we can generate the most probable spectrogram for each β.

2) Simulation Conditions: To analyze the net extrapolation ability, we simulate the spectrogram restoration task. In this simulation, we generated random i.i.d. values, which obey the corresponding generation model (32) for each β, as the observed data matrix $\boldsymbol{Y}$. We compared $\beta=0$, 1, 2, 3, 4 and $\beta_{\mathrm{R}}=0$, 1, 2, 3, and we used the same divergence β in the training and separation processes. The size of this data matrix was set to $\Omega=5000$ and $T=200$. We set the

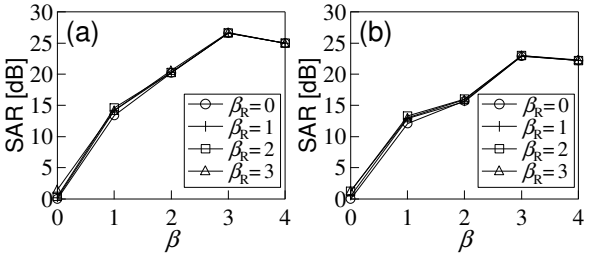

Fig. 5. Extrapolation abilities for (a) 75%-binary-masked data and (b) 98%-binary-masked data.

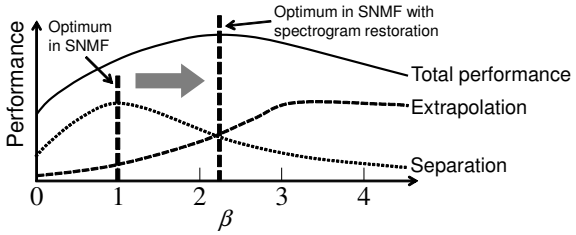

Fig. 6. Trade-off between separation and extrapolation abilities. Overall performance is highest when $\beta > 1$.

parameters of each p.d.f. to $\vartheta_1 = 1$, $\vartheta_2 = 5$, $\vartheta_3 = 10$, $\vartheta_4 = 50$, $\vartheta_5 = 2$, and $M = 15$. These parameters are determined so as to generate the nonnegative random i.i.d. values that obey each corresponding generation model. Note that the parameters θ_1–θ_5 simply determine the scales of the input random variables, and basically can be set to arbitrary value without loss of generality. In addition, we used two types of data-missing patterns $\boldsymbol{I}$, in which 75% or 98% of the grids were missing in a uniform manner, and the missing data $\boldsymbol{I} \circ \boldsymbol{Y}$ imitated the binary-masking procedure. The supervised bases $\boldsymbol{F}$ were obtained by training using the same data matrix $\boldsymbol{Y}$, namely, $\boldsymbol{Y}_{\rm target} = \boldsymbol{Y}$ in (7) and (8). The number of supervised bases, K, was 100, which is the half size of T, and the number of other bases, L, was 30. Therefore, the task was to reconstruct original $\boldsymbol{Y}$ from the observations with missing data, $\boldsymbol{I} \circ \boldsymbol{Y}$, using the trained bases.

3) Simulation Results and Discussion: We used sources-to-artifacts ratio (SAR) defined in [17] as the accuracy of the extrapolation,where SAR indicates the absence of artificial distortion. Figure 5 shows the SAR result for each divergence and regularization. From this result, it is confirmed that a higher β provides better basis extrapolation regardless of the type of regularization ($\beta_{\rm R}$). In NMF decomposition, if we set β to a large value, the trained bases tend to become anti-sparse (smooth). In contrast, if β is close to zero, the trained bases become more sparsity-aware, and this property is suitable for normal NMF-based music source separation because of the inherent sparsity of music spectrograms (e.g., $\beta = 1$ is recommended in [14], [15]). However, for basis extrapolation, sparse bases are *not* suitable because it is difficult to extrapolate them only from the observable data. Therefore, we speculate that the optimal divergence in SNMF with spectrogram restoration, which attempts to fit the trained bases using spectral components except for chasms, is shifted to $\beta > 1$ rather than KL-divergence ($\beta = 1$) because of the trade-off between separation and extrapolation abilities, as illustrated in Fig. 6. This issue will be confirmed experimentally in the next section.

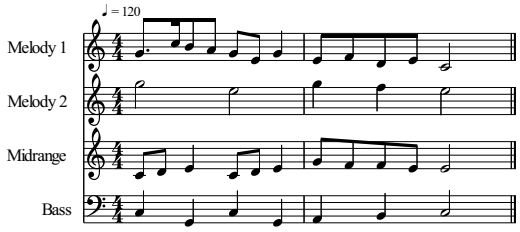

Fig. 7. Scores of each part.

TABLE I
COMPOSITIONS OF MUSICAL INSTRUMENTS

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

C. Comparison Between Proposed Hybrid Method and Conventional Methods

1) Experimental Conditions: We conducted objective evaluation to confirm the effectiveness of the proposed hybrid method described in the previous section. In this experiment, we compare the separation performance of five methods, namely simple directional clustering [10], multichannel NMF [9], simple SNMF [14], naive hybrid method described in Sect. III-A1, and the proposed hybrid method including SNMF with spectrogram restoration after directional clustering, in terms of their ability to separate music artificial and real-recorded signals. Also, we compared some evaluation scores with various β and $\beta_{\rm R}$ for SNMF, naive hybrid method, and the proposed hybrid method by setting five divergences and three regularizations, namely, $\beta = 0$, 1, 2, 3, 4 and $\beta_{\rm R} = 0$, 1, 2. We used the same divergence (β) in the training and separation processes for SNMF, naive hybrid method, and proposed hybrid method. 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. 7) with three compositions (C1–C3) of instruments shown in Table I. These signals were artificially generated by a MIDI synthesizer. In particular, these stereo signals were mixed down to a monaural format only when we evaluate the separation accuracy of SNMF because SNMF is a separation method for a monaural input signal.

In the artificial signal case, the observed signals $\boldsymbol{Y}$ were produced by mixing four sources with the same power. The observed signal contained one source in the left and right directions and two sources in the center direction based on a sine law (see Fig. 8 (a)). The target instrument is always located in the center direction along with another interfering instrument, and we prepared two patterns in which the left and right sources are located at $45°$. In addition, we used the same MIDI sounds of the target instruments as supervision for a priori training. The training sounds contained two octave notes that cover all the notes of the target signal in the

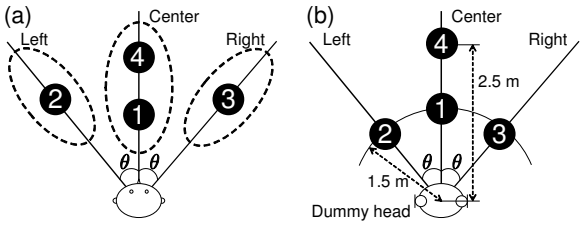

Fig. 8. Location of four sources with sine law used in (a) artificial signal and (b) real-recorded signal cases. Numbered black circles represent locations of instruments in stereo format.

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 for the training and separation were 500. Moreover, the number of clusters used in directional clustering was 3, the number of a priori bases, K, was 100, and the number of bases for matrix $\boldsymbol{H}$, L, was 30. The weighting 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). A geometry of the loudspeaker and binaural microphone is shown in Fig. 8 (b), where $\theta = 45°$. The target source and the supervision sound are always located in No.1 position in Fig. 8 (b). The observed signal $\boldsymbol{Y}$ was produced by mixing these recorded signals as the same power. Other conditions were the same as those of the artificial signal case.

2) Experimental Results: We used the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and SAR defined in [17]. SDR indicates the quality of the separated target sound, and SIR indicates the degree of separation between the target and other sounds. Therefore, SDR indicates the total evaluation score that involves SIR and SAR.

Figure 9 shows the average SDR, SIR, and SAR of the proposed hybrid method and the other methods for each divergence (β) and each regularization (β_{R}) in the artificial signal case, where the four instruments are shuffled with 12 combinations in each of compositions C1–C3. Therefore, these results are the averages of 36 input signal patterns. Also, Fig. 10 shows the average SDR, SIR, and SAR in the real-recorded signal case. From the SDRs in Figs. 9 and 10, we can confirm that directional clustering does not have sufficient performance because this method cannot discriminate the sources in the same direction. Multichannel NMF also cannot achieve the sufficient separation because this method strongly depends on the initial value and lack robustness. In contrast, the methods using SNMF can give better results and the proposed hybrid method using SNMF with spectrogram restoration outperforms all other methods in both artificial and real-recorded signal cases. The naive hybrid method is inferior to SNMF when $\beta \leq 1$ whereas this hybrid method utilizes both directional clustering and SNMF. This is because

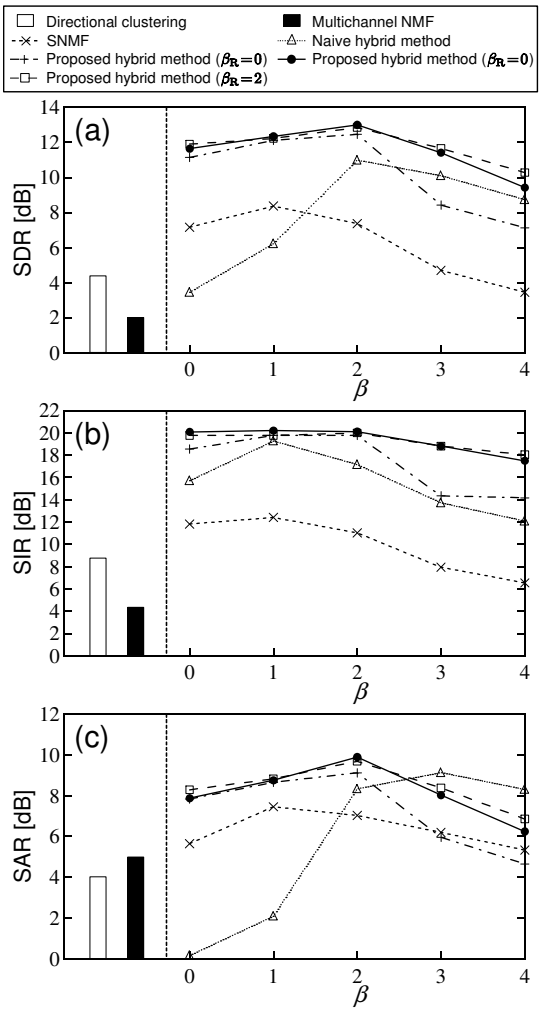

Fig. 9. Average scores in artificial signal case: (a) shows SDR, (b) shows SIR, and (c) shows SAR for conventional and proposed methods.

the naive hybrid method is affected by the spectral chasms and cannot reconstruct such lost components. Furthermore, we can confirm that the EUC-distance-based cost function ($\beta = 2$) is an optimal divergence for the proposed hybrid method, whereas KL-divergence ($\beta = 1$) is the best divergence even for conventional SNMF [14], [15]. This marked shift of the optimal divergence in SNMF with spectrogram restoration is due to the trade-off between the separation and extrapolation abilities, as predicted in Sect. III-B. In addition, the regularization with KL-divergence ($\beta_{\mathrm{R}} = 1$) is slightly better than the other divergences but the difference is not significant, except for the case of $\beta_{\mathrm{R}} = 0$.

IV. SNMF with Spectrogram Restoration Based on Multi-Divergence

A. Divergence Dependency on Local Chasms Condition

In the previous section, we revealed the mechanism of optimal divergence shift in the SNMF methods. This divergence shift is due to the trade-off between separation and extrapolation abilities. The optimal divergence for SNMF with spectrogram restoration depends on the rate of spectral

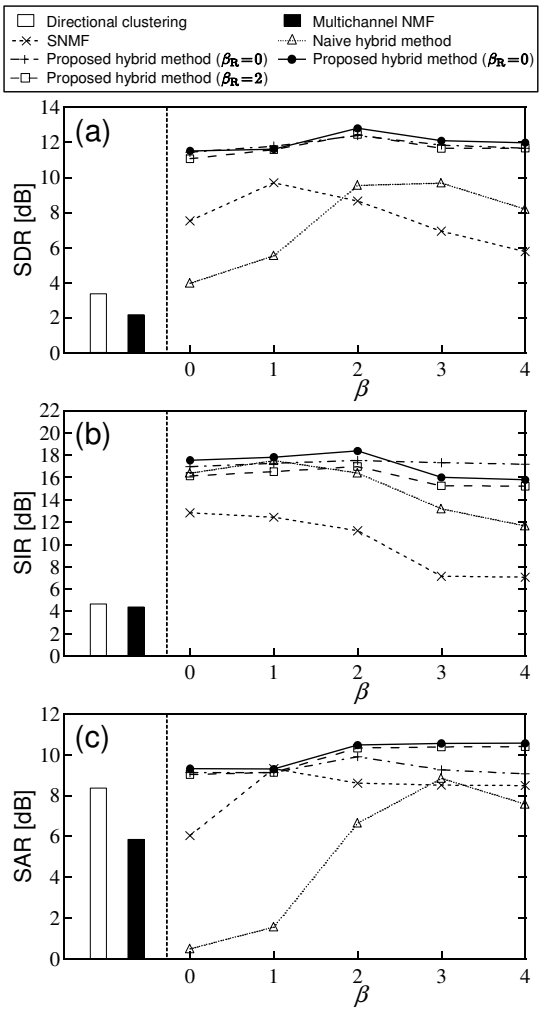

Fig. 10. Average scores in real-recorded signal case: (a) shows SDR, (b) shows SIR, and (c) shows SAR for conventional and proposed methods.

chasms in each time frame of the spectrogram obtained by preceding directional clustering. Therefore, the optimal divergence temporally fluctuates because the spatial condition is not consistent in the general music signal, and the divergence of SNMF should be changed in each time frame automatically. To solve this problem, in this section, we propose a new scheme for frame-wise divergence selection to separate the target signal using optimal divergence.

If there are many chasms in a frame of the binary-masked spectrogram, SNMF is preferred to have high extrapolation ability. In contrast, if the rate of chasms is low value, the separation ability is required rather than the extrapolation. Therefore, it is expected that EUC-distance should be used in the frames that have many chasms, and KL-divergence should be used in the other frames. To improve total separation performance of SNMF with spectrogram restoration for any types of input signals, we introduce a multi-divergence-based cost function as described in the next section.

B. Cost Function and Update Rules

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 with spectrogram restoration 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 spectrogram restoration that includes a multi-divergence-based cost function covered onto the whole input matrix (see Fig. 11). The proposed cost function $\mathcal{J}_{\mathrm{m}}$ is defined as

$$\mathcal{J}_{\mathrm{m}} = \textstyle\sum_t \mathcal{J}_t, \tag{34}$$

$$\mathcal{J}_t = \begin{cases} \sum_\omega i_{\omega t} \mathcal{D}_{\beta=2}(y_{\omega t} \| s^{(\mathrm{E})}_{\omega t}) \\ \quad + \lambda^{(\mathrm{E})} \sum_\omega \overline{i_{\omega t}} \mathcal{D}_{\beta_{\mathrm{R}}}(0 \| \sum_k f^{(\mathrm{E})}_{\omega k} g_{kt}) \\ \quad + \mu^{(\mathrm{E})} \| \boldsymbol{F}^{(\mathrm{E})\mathrm{T}} \boldsymbol{H} \|^2_{\mathrm{Fr}} & (r_t \geq \tau) \\ \sum_\omega i_{\omega t} \mathcal{D}_{\beta=1}(y_{\omega t} \| s^{(\mathrm{K})}_{\omega t}) \\ \quad + \lambda^{(\mathrm{K})} \sum_\omega \overline{i_{\omega t}} \mathcal{D}_{\beta_{\mathrm{R}}}(0 \| \sum_k f^{(\mathrm{K})}_{\omega k} g_{kt}) \\ \quad + \mu^{(\mathrm{K})} \| \boldsymbol{F}^{(\mathrm{K})\mathrm{T}} \boldsymbol{H} \|^2_{\mathrm{Fr}} & (r_t < \tau) \end{cases}, \tag{35}$$

$$s^{(*)}_{\omega t} = \textstyle\sum_k f^{(*)}_{\omega k} g_{kt} + \sum_n h_{\omega n} u_{nt}, \tag{36}$$

$$r_t = \textstyle\sum_\omega \overline{i_{\omega t}} / \Omega, \tag{37}$$

where $\boldsymbol{F}^{(\mathrm{K})} (\in \mathbb{R}^{\Omega \times K}_{\geq 0})$ and $\boldsymbol{F}^{(\mathrm{E})} (\in \mathbb{R}^{\Omega \times K}_{\geq 0})$ are the supervised basis matrices trained in advance using KL-divergence-based NMF and EUC-distance-based NMF, respectively. Also, $f^{(\mathrm{K})}_{\omega k}$ and $f^{(\mathrm{E})}_{\omega k}$ are the entries of $\boldsymbol{F}^{(\mathrm{K})}$ and $\boldsymbol{F}^{(\mathrm{E})}$, respectively, $\mu^{(*)}$ and $\lambda^{(*)}$ are the weighting parameters for each term, and $* = \{\mathrm{K}, \mathrm{E}\}$. 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. Similarly to Sect. III-A2, we can derive the update rules based on (34) by the auxiliary function approach as follows:

$$g_{kt} \leftarrow \begin{cases} g_{kt} \cdot \dfrac{\sum_\omega i_{\omega t} y_{\omega t} f^{(\mathrm{E})}_{\omega k}}{\sum_\omega i_{\omega t} f^{(\mathrm{E})}_{\omega k} s^{(\mathrm{E})}_{\omega t} + \lambda^{(\mathrm{E})} R^{(\mathrm{E})}_G} & (r_t \geq \tau) \\ g_{kt} \cdot \dfrac{\sum_\omega i_{\omega t} y_{\omega t} f^{(\mathrm{K})}_{\omega k} s^{(\mathrm{K})-1}_{\omega t}}{\sum_\omega i_{\omega t} f^{(\mathrm{K})}_{\omega k} + \lambda^{(\mathrm{K})} R^{(\mathrm{K})}_G} & (r_t < \tau) \end{cases}, \tag{38}$$

$$h_{\omega l} \leftarrow h_{\omega l} \cdot \frac{\sum_t i_{\omega t} y_{\omega t} u_{lt} N_{\omega t}}{\sum_t i_{\omega t} u_{lt} D_{\omega t} + P_{\omega l}}, \tag{39}$$

$$u_{lt} \leftarrow \begin{cases} u_{lt} \cdot \dfrac{\sum_\omega i_{\omega t} y_{\omega t} h_{\omega l}}{\sum_\omega i_{\omega t} h_{\omega l} s^{(\mathrm{E})}_{\omega t}} & (r_t \geq \tau) \\ u_{lt} \cdot \dfrac{\sum_\omega i_{\omega t} y_{\omega t} h_{\omega l} s^{(\mathrm{E})-1}_{\omega t}}{\sum_\omega i_{\omega t} h_{\omega l}} & (r_t < \tau) \end{cases}, \tag{40}$$

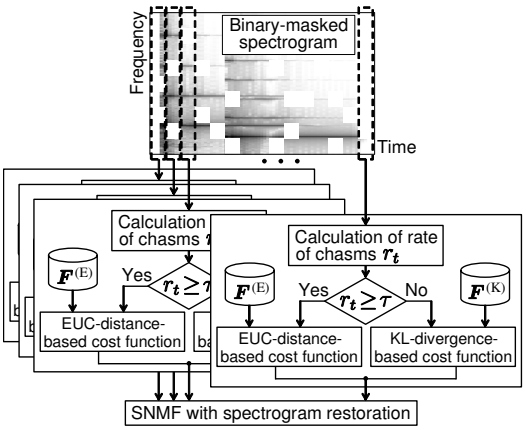

Fig. 11. Multi-divergence algorithm of proposed method.

where $R_G^{(*)}$, $N_{\omega t}$, $D_{\omega t}$, and $P_{\omega l}$ are given by

$$R_G^{(*)} = \sum_\omega \overline{i_{\omega t}} f_{\omega k}^{(*)} \left(\sum_{k'} f_{\omega k'}^{(*)} g_{k' t} \right)^{\beta_{\mathrm{R}}-1}, \tag{41}$$

$$N_{\omega t} = \begin{cases} 1 & (r_t \geq \tau) \\ s_{\omega t}^{(\mathrm{K})-1} & (r_t < \tau) \end{cases}, \tag{42}$$

$$D_{\omega t} = \begin{cases} s_{\omega t}^{(\mathrm{E})} & (r_t \geq \tau) \\ 1 & (r_t < \tau) \end{cases}, \tag{43}$$

$$P_{\omega l} = \begin{cases} \mu^{(\mathrm{E})} \sum_k f_{\omega k}^{(\mathrm{E})} \sum_{\omega'} f_{\omega' k}^{(\mathrm{E})} h_{\omega' l} & (r_t \geq \tau) \\ \mu^{(\mathrm{K})} \sum_k f_{\omega k}^{(\mathrm{K})} \sum_{\omega'} f_{\omega' k}^{(\mathrm{K})} h_{\omega' l} & (r_t < \tau) \end{cases}. \tag{44}$$

C. Evaluation Experiment

1) Experimental Conditions: To confirm the effectiveness of the proposed algorithm, we compared six methods, namely, SNMF based on KL-divergence and EUC-distance [14], simple directional clustering [10], multichannel NMF [9], the conventional hybrid method based on KL-divergence and EUC-distance, the online hybrid method described in Sect. IV-B, and the proposed hybrid method that uses multi-divergence.

In this experiment, similarly to Sect. III-C1, we produced the artificial and real-recorded stereo signals containing four melody parts (depicted in Fig. 12) with three compositions (C1–C3) of instruments shown in Table I. These stereo signals were mixed down to a monaural format only when we evaluate the separation accuracy of SNMF. In addition, we prepared four spatially different dataset patterns of the observed signals, SP1–SP4, as shown in Table II. In the hybrid method, many chasms were produced by directional clustering in the measures where $\theta = 45^\circ$ compared with those of $\theta = 0^\circ$. Therefore, we can expect that EUC-distance-based hybrid method is suitable for SP4 rather than the dataset of SP1. The threshold value τ, was set to 20%, and the type of regularization was $\beta_{\mathrm{R}} = 1$. The other experimental conditions were the same as those in Sect. III-C1.

2) Experimental Results: Figures 13 and 14 show the average SDR, SIR, and SAR scores for each method and each dataset pattern, where these results are the averages of 36 input signal patterns, similarly to Sect. III-C1. The SDR scores of SNMF are the same for any datasets because the input signals for SNMF are mixed down to a monaural format.

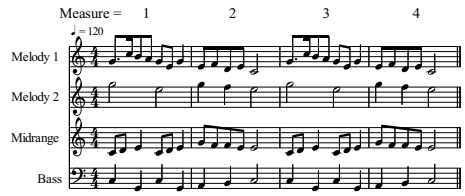

Fig. 12. Scores of each part. The observed signal consists of four measures.

TABLE II
SPATIAL CONDITIONS OF EACH DATASET

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

From this result, 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 ($\theta = 45^\circ$), which yields many chasms, and high extrapolation ability is required. In addition, 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.

V. CONCLUSION

In this paper, first, we proposed a new multichannel signal separation method, i.e., a hybrid method that concatenates SNMF with spectrogram restoration after directional clustering. The proposed SNMF with spectrogram restoration attempts both target signal separation and reconstruction of the lost target components, which are generated by preceding binary masking performed in directional clustering.

Secondly, from the theoretical analysis, it was revealed that the optimal divergence in SNMF with spectrogram restoration is shifted to an anti-sparse divergence rather than KL-divergence. This was due to the fact that there exists the trade-off between separation and extrapolation abilities in SNMF. Evaluation experiment of the separation using artificial and real-recorded music signals showed the effectiveness of the proposed hybrid method.

Finally, on the basis of this finding, we also proposed an improved hybrid method based on multi-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 showed that our proposed method can always achieve high separation accuracy under all spatial conditions.

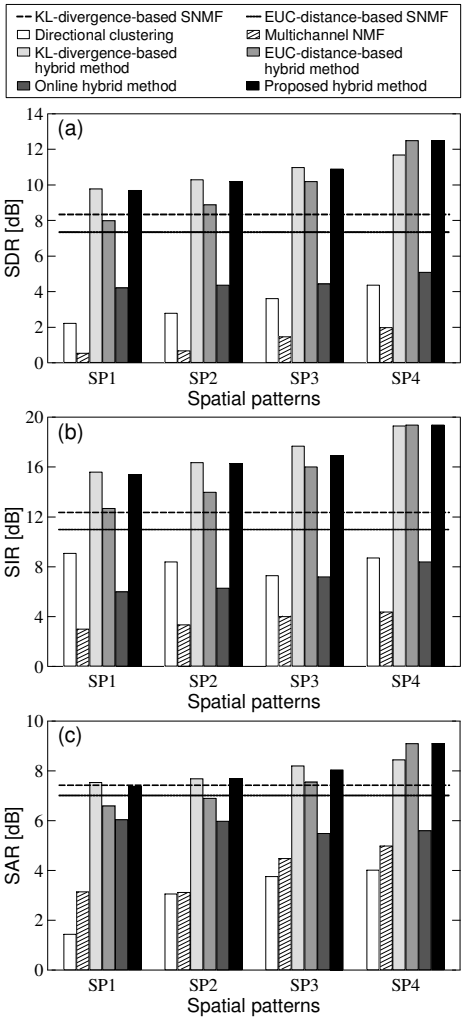

Fig. 13. Average scores of each method and each spatial condition in artificial signal case: (a) shows SDR, (b) shows SIR, and (c) shows SAR.

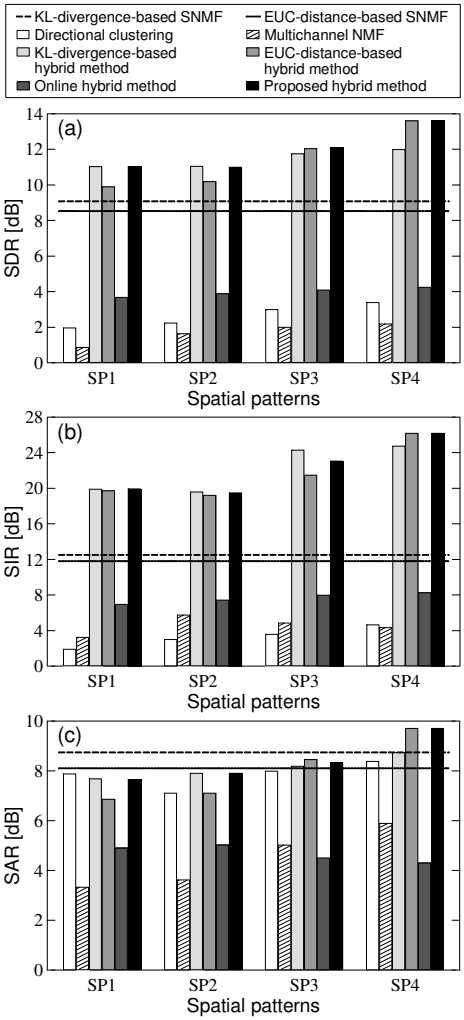

Fig. 14. Average scores of each method and each spatial condition in real-recorded signal case: (a) shows SDR, (b) shows SIR, and (c) shows SAR.

REFERENCES

[1] D. Kitamura, H. Saruwatari, H. Kameoka, K. Kondo, Y. Takahashi, S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE Trans. ASLP*, 2014 (under review).

[2] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," *Proc. EUSIPCO*, 2008.

[3] A. Mesaros, T. Virtanen, A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," *Proc. ISMIR*, pp.375–378, 2007.

[4] 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.

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

[6] 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. ICASSP*, pp.5365–5368, 2012.

[7] 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.

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

[9] H. Sawada, H. Kameoka, S. Araki, N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," *Proc. ICASSP*, pp.261–264, 2012.

[10] 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.

[11] Y. Li, D. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Communication*, vol.51, no.3, pp.230–239, 2009.

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

[13] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, S. Sagayama, "Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence," *Proc. MLSP*, pp.283–288, 2010.

[14] 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, pp.1113–1118, 2014.

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

[16] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol.2009, p.1–17, 2009.

[17] E. Vincent, R. Gribonval, C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol.14, no.4, pp.1462–1469, 2006.