Multichannel Signal Separation Combining Directional Clustering and Nonnegative Matrix Factorization with Spectrogram Restoration

Daichi Kitamura, *Member, IEEE*, Hiroshi Saruwatari, *Member, IEEE*, Hirokazu Kameoka, *Member, IEEE*, Yu Takahashi, *Member, IEEE*, Kazunobu Kondo, *Member, IEEE*, and Satoshi Nakamura, *Senior Member, IEEE*

***Abstract*—In this paper, to address problems in multichannel music signal separation, we propose a new hybrid method that combines directional clustering and advanced nonnegative matrix factorization (NMF). The aims of multichannel music signal separation technology is to extract a specific target signal from observed multichannel signals that contain multiple instrumental sounds. In previous studies, various methods using NMF have been proposed, but many problems remain including poor separation accuracy and lack of robustness. To solve these problems, we propose a new supervised NMF (SNMF) with spectrogram restoration and a hybrid method that concatenates the proposed SNMF after directional clustering. Via the 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 experimentally reveal the trade-off between separation and extrapolation abilities and propose a new scheme for adaptive divergence, where the 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.**

***Index Terms*—Multichannel signal separation, music signal processing, nonnegative matrix factorization (NMF), spectrogram restoration.**

I. Introduction

MUSIC signal separation technologies have attracted considerable interest and been intensively studied [1], [2] 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 at least two. As a means of addressing underdetermined signal separation, in recent years, nonnegative matrix factorization (NMF) [3], 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 [4]–[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]–[9]. This method includes a priori training, which requires some sound samples of a target instrument, and separates 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 [10], [11]. 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 suffer from poor separation accuracy and lack robustness. For multichannel signal separation, directional clustering has also been proposed as an unsupervised method [12], [13]. 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. Furthermore, the extracted signal is likely to be distorted because some target components may be lost by the effect of binary masking in the directional clustering.

To cope with these problems, in this paper, we propose a new SNMF with spectrogram restoration and a hybrid method that concatenates the proposed SNMF after directional clustering. This approach can reconstruct lost target components, which are dispersedly generated by directional clustering, from only the observable valid components using supervised bases. Such

Manuscript received May 22, 2014; revised October 06, 2014; accepted January 20, 2015. Date of publication February 06, 2015; date of current version March 06, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. DeLiang Wang.

D. Kitamura is with the Department of Informatics, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies, Tokyo 101–8430, Japan (e-mail: d-kitamura@nii.ac.jp).

H. Saruwatari and H. Kameoka are with the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113–8656, Japan (e-mail: hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp; kameoka@hil.t.u-tokyo.ac.jp).

Y. Takahashi and K. Kondo are with the Research and Development 1, Yamaha Corporation, Shizuoka 438–0192, Japan (e-mail: yu.takahashi@music.yamaha.com; kazunobu.kondo@music.yamaha.com).

S. Nakamura is with the Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630–0192, Japan (e-mail: s-nakamura@is.naist.jp).

Digital Object Identifier 10.1109/TASLP.2015.2401425

reconstruction with supervised bases can be considered as *basis extrapolation*. In [14], bandwidth expansion with supervised basis extrapolation was proposed. However, this method only attempts to predict unseen high-frequency components. In contrast, our proposed method reconstructs dispersedly lost components in parallel with source separation. Via the extrapolation of supervised spectral bases, SNMF with spectrogram restoration attempts both target signal separation and reconstruction of the lost target components, which are generated by the preceding binary masking performed in directional clustering.

Next, we provide an experimental analysis of basis extrapolation ability and reveal the mechanism of the marked shift of the optimal divergence in SNMF with spectrogram restoration and the trade-off between separation and extrapolation abilities. An evaluation experiment of the separation using artificial and real-recorded music signals shows the effectiveness of the proposed hybrid method.

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

The rest of this paper is organized as follows. In Section II, conventional methods for single-channel and multichannel signal separation are described. In Section III, we propose a new SNMF with spectrogram restoration and a hybrid method and experimentally reveal the trade-off between separation and extrapolation abilities. In Section IV, an improved method based on adaptive divergence is presented. Following a discussion on the results of the experiments, we present our conclusions in Section V.

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 $\boldsymbol{X}$, respectively, and D is the number of bases of $\boldsymbol{V}$. Fig. 1 depicts the decomposition model of NMF, where the number of bases D equals two. In this Fig., 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

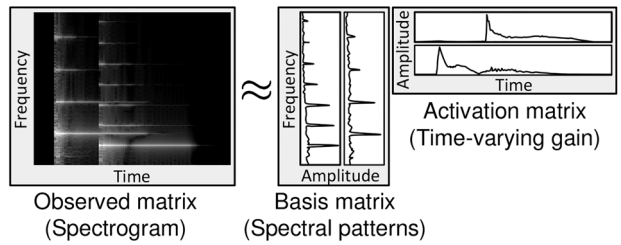

Fig. 1. Decomposition model of simple NMF.

$\boldsymbol{X}$ and $\boldsymbol{V}\boldsymbol{W}$. The following equation represents the cost function of NMF:

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

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 [15] in the cost function:

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

where $\boldsymbol{A}(\in \mathbb{R}^{I\times J})$ and $\boldsymbol{B}(\in \mathbb{R}^{I\times J})$ are matrices whose entries are $a_{i,j}$ and $b_{i,j}$, 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).

The multiplicative update rules for $\boldsymbol{V}$ and $\boldsymbol{W}$ that minimize the cost function based on β-divergence are given by [16]

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

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

where $x_{m,n}$, $v_{m,d}$, and $w_{d,n}$ 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 has been theoretically proven for all real values of β [16].

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 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]–[9]. This 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 the amplitude spectrogram of a specific instrumental signal used 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 used for the supervision of the target instrumental spectrum.

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

$$\boldsymbol{Y} \simeq \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}, \tag{8}$$

where $\boldsymbol{Y} (\in \mathbb{R}_{\geq 0}^{\Omega \times T})$ is the observed spectrogram, $\boldsymbol{G} (\in \mathbb{R}_{\geq 0}^{K \times T})$ is the activation matrix that corresponds to $\boldsymbol{F}$, $\boldsymbol{H} (\in \mathbb{R}_{\geq 0}^{\Omega \times L})$ is 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, T is the number of frames of the observed signal and L is the number of bases of $\boldsymbol{H}$. Strictly speaking, some papers call this method semi-supervised NMF to discriminate between the words "semi-supervised" (only the target sound is trained) and "fully supervised" (the target and interference sounds are trained). However, we simply describe this method as "supervised" in this paper because we do not intend to compare semi-supervised and fully supervised cases, as reported in other papers. In SNMF, the matrices $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ are optimized under the condition that $\boldsymbol{F}$ is known in advance. Hence, $\boldsymbol{F}\boldsymbol{G}$ ideally represents the target instrumental component and $\boldsymbol{H}\boldsymbol{U}$ represents other interfering components after the decomposition. The cost function for (8) is defined as

$$\mathcal{J}_{\mathrm{SNMF}} = \mathcal{D}_{\beta} \left(\boldsymbol{Y} \| \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U} \right). \tag{9}$$

Also, the update rules for (9) are given by

$$h_{\omega,l} \leftarrow h_{\omega,l} \left(\frac{\sum_t y_{\omega,t} u_{l,t} z_{\omega,t}^{\beta-2}}{\sum_t u_{l,t} z_{\omega,t}^{\beta-1}} \right)^{\varphi(\beta)}, \tag{10}$$

$$g_{k,t} \leftarrow g_{k,t} \left(\frac{\sum_\omega f_{\omega,k} y_{\omega,t} z_{\omega,t}^{\beta-2}}{\sum_\omega f_{\omega,k} z_{\omega,t}^{\beta-1}} \right)^{\varphi(\beta)}, \tag{11}$$

$$u_{l,t} \leftarrow u_{l,t} \left(\frac{\sum_\omega h_{\omega,l} y_{\omega,t} z_{\omega,t}^{\beta-2}}{\sum_\omega h_{\omega,l} z_{\omega,t}^{\beta-1}} \right)^{\varphi(\beta)}, \tag{12}$$

where $y_{\omega,t}$, $f_{\omega,k}$, $g_{k,t}$, $h_{\omega,l}$, and $u_{l,t}$ are the nonnegative entries of the matrices $\boldsymbol{Y}$, $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$, respectively, and

$$z_{\omega,t} = \sum_k f_{\omega,k} g_{k,t} + \sum_l h_{\omega,l} u_{l,t}. \tag{13}$$

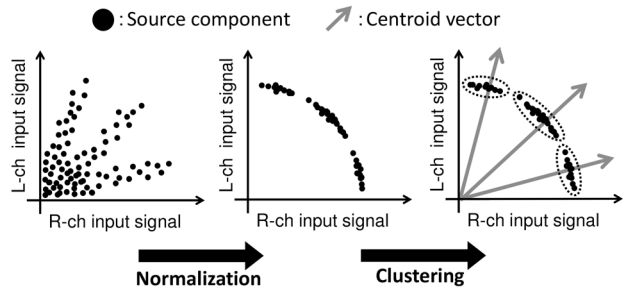

Fig. 2. Configuration of directional clustering.

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 of the multichannel signal have also been proposed as unsupervised separation techniques [12], [13]. In this paper, we only focus on the gain-based directional clustering method, which is a simple version of the technique in [12]. Fig. 2 shows the configuration of directional clustering. First, the time-frequency components of a stereo mixed signal $\boldsymbol{y}_{\omega,t} = [y_{\omega,t}^{(L)} y_{\omega,t}^{(R)}]$ are represented in a two-dimensional space having the amplitude of each channel as the coordinate axes, where $y_{\omega,t}^{(L)}$ and $y_{\omega,t}^{(R)}$ are the amplitudes of the left and right channels, respectively. Next, these components are normalized over the unit circle to make apparent clusters, which correspond to each directional component. Finally, these clusters are separated by the k-means clustering method. Therefore, this method is equivalent to the quantization of 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 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 components is degraded by time-frequency binary masking. Fig. 3 shows an example of the spectrum of a signal separated by directional clustering. The obtained spectrum has many chasms owing to the 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 [10], [11]. The algorithms used in this method employ a Hermitian positive definite matrix that models the spatial property of each NMF basis and each frequency bin. Therefore,

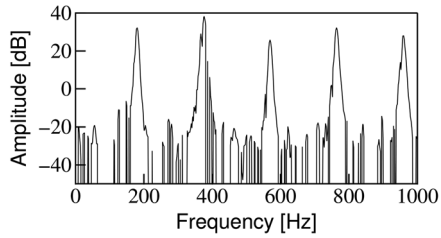

Fig. 3. Example of spectrum of signal separated by directional clustering.

multichannel NMF utilizes a frequency-wise transfer function between the 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 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 a few instruments. Moreover, the residual interfering signal in the same direction can be removed by SNMF.

However, such naive hybrid method has a problem that the extracted signal may suffer from the generation of considerable distortion. This is because the spectrogram obtained from directional clustering has many spectral chasms owing to the binary masking procedure. These spectral losses may deteriorate the separation performance 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 [17].

Fig. 4 shows the signal flow in 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. 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 via supervised basis extrapolation. Furthermore, a soft directional mask, which employs probabilities instead of binary indexes, can also be applied to the proposed hybrid method (see Appendix A).

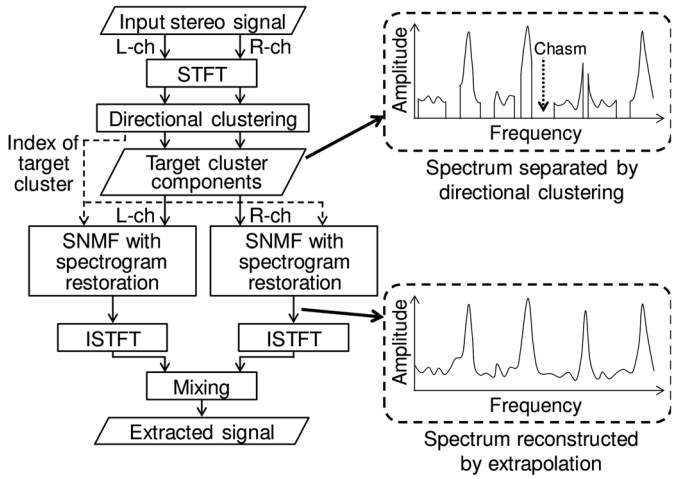

Fig. 4. Signal flow of proposed hybrid method; SNMF with spectrogram restoration is performed after directional clustering.

The proposed method requires directional information of the target signal, namely, we have to know which directional cluster includes the target signal. In this paper, the target signal is always located in the center direction, and such a priori information for the direction is given. However, when the target signal is located in an other direction, we can apply the proposed hybrid method in the same manner. In addition, even if the target direction is unknown, we can obtain the separated signal by applying SNMF with spectrogram restoration to all the directions (clusters) and choosing the result with the highest quality.

To illustrate the separation mechanism step by step, Fig. 5(a) shows the direction of arrival (D.O.A.) histogram of each source (shaded with various patterns to distinguish the sources) in the stereo signal, (b) shows the separated components that are clustered around the center direction after directional clustering, and (c) shows the separated target component obtained by SNMF with spectrogram restoration. In Fig. 5(a), the source components are distributed in all directions with some overlapping. This is because the sound sources are received with the room reverberation. After directional clustering (Fig. 5(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. The lost tail of the center sources corresponds to the binary-masked points in the time-frequency domain, and the leaked tails in the center cluster are the components of left- and right-side interference sources, which are not masked in directional clustering. After SNMF with spectrogram restoration, the proposed algorithm restores the lost components by supervised basis extrapolation (Fig. 5(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 all zero, a large extrapolation error may occur. Then, incorrect bases are chosen and fitted to a small number of time-frequency points by incorrectly modifying the activation matrix $\boldsymbol{G}$. In the worst case, the activation matrix $\boldsymbol{G}$ contains very large values at a specific time. For example, when only one grid point is observed and the other points are masked in a frame, this frame is able to be extrapolated with any type of supervised bases.

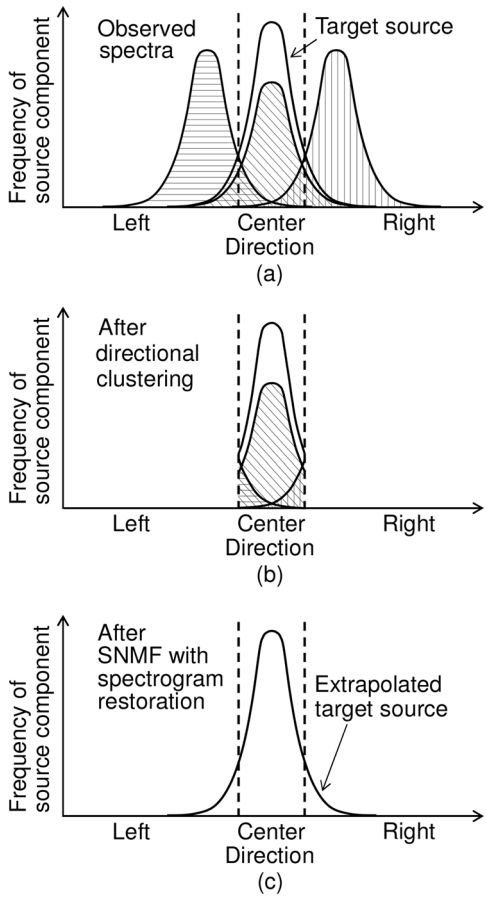

Fig. 5. Directional source distribution of (a) observed stereo signal, (b) separated components in center cluster, and (c) component separated and extrapolated by spectrogram restoration.

If such an observed grid point has large value but the chosen basis has a spectral valley at this grid point, a large gain of $\boldsymbol{G}$ is generated for the chosen basis; this leads to unexpected spectral peaks outside the observed grid point. Such an extrapolation error generates very loud and unnatural sounds in the waveform domain. To avoid this, we should add a new penalty term [18] in the cost function, as described in the next section.

2) Cost Function: We define the cost function of SNMF with spectrogram restoration using β-divergence. 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 indicates whether or not each grid point 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

$$\begin{aligned}\mathcal{J}(\Theta) = &\sum_{\omega,t} i_{\omega,t}\mathcal{D}_{\beta}\left(y_{\omega,t}\|\textstyle\sum_k f_{\omega,k}g_{k,t} + \sum_l h_{\omega,l}u_{l,t}\right)\\ &+ \lambda\sum_{\omega,t}\overline{i_{\omega,t}}\mathcal{D}_{\beta_{\mathrm{R}}}\left(0\|\textstyle\sum_k f_{\omega,k}g_{k,t}\right) + \mu\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\mathrm{F}}^2,\end{aligned} \tag{14}$$

where $\Theta = \{\boldsymbol{G}, \boldsymbol{H}, \boldsymbol{U}\}$ 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, $\|\cdot\|_{\mathrm{F}}$ is a Frobenius norm, 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 spectral grid point 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 $\sum_k f_{\omega,k}g_{k,t}$. 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 [18]. In other words, this penalty term regulates the extrapolation. As another means of avoiding the extrapolation error, some people may guess that a simple sparse regularization for the activation $\boldsymbol{G}$ can also be introduced instead of the proposed regularization. This issue will be discussed in Appendix B. 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 the separation performance [19].

3) Auxiliary Function Technique: The update rules of NMF are usually derived by the *auxiliary function technique*, which is an extension of the expectation-maximization algorithm. To explain this technique, let us consider a general optimization problem of finding an optimum parameter vector $\Theta = \Theta^{\dagger}$ that satisfies

$$\Theta^{\dagger} = \arg\min_{\Theta} \mathcal{F}(\Theta), \tag{15}$$

where F is a cost function. In the auxiliary function technique, we have to find an auxiliary function $F^{+}(\Theta, \hat{\Theta})$ satisfying

$$\mathcal{F}(\Theta) = \min_{\hat{\Theta}} \mathcal{F}^{+}(\Theta, \hat{\Theta}), \tag{16}$$

where $\hat{\Theta}$ are called auxiliary variables. Then, instead of directly minimizing the cost function $\mathcal{F}(\Theta)$, the auxiliary function $\mathcal{F}^{+}(\Theta, \hat{\Theta})$ is minimized in terms of Θ and $\hat{\Theta}$, alternately. The iterative update rules are obtained as

$$\hat{\Theta}^{\mathrm{new}} = \arg\min_{\hat{\Theta}} \mathcal{F}^{+}(\Theta, \hat{\Theta}), \tag{17}$$

$$\Theta^{\mathrm{new}} = \arg\min_{\Theta} \mathcal{F}^{+}(\Theta, \hat{\Theta}^{\mathrm{new}}). \tag{18}$$

In these updates, a monotonic decrease in $F(\Theta)$ is guaranteed. In addition, the update rules of auxiliary variables in an NMF-based method can usually be written in a closed form, and we can obtain efficient update rules for NMF variables.

4) Derivation of Update Rules: Similarly to in [16], we derive the update rules based on β-divergence using an auxiliary function technique. Here, we rewrite the cost function (14) using β-divergence as

$$\mathcal{J}(\Theta) = \mathcal{J}_1 + \lambda\mathcal{J}_2 + \mu\mathcal{J}_3, \tag{19}$$

$$\mathcal{J}_1 = \sum_{\omega,t} i_{\omega,t}\left(\frac{z_{\omega,t}^{\beta}}{\beta} - \frac{y_{\omega,t}z_{\omega,t}^{\beta-1}}{\beta-1}\right), \tag{20}$$

$$\mathcal{J}_2 = \sum_{\omega,t} \overline{i_{\omega,t}}\frac{(\sum_k f_{\omega,k}g_{k,t})^{\beta_{\mathrm{R}}}}{\beta_{\mathrm{R}}}, \tag{21}$$

$$\mathcal{J}_3 = \sum_{k,l}\left(\sum_{\omega} f_{\omega,k}h_{\omega,l}\right)^2, \tag{22}$$

where constant terms are omitted.

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 $\alpha_{\omega,t,k} \geq 0$, $\gamma_{\omega,t,l} \geq 0$, $\eta_1 \geq 0$, $\eta_2 \geq 0$, and $\sigma_{\omega,t}$ that satisfy $\sum_k \alpha_{\omega,t,k} = 1$, $\sum_l \gamma_{\omega,t,l} = 1$, and $\eta_1 + \eta_2 = 1$ as

$$\mathcal{J}_1 \leq \mathcal{J}_1^+ = \sum_{\omega,t} i_{\omega,t} \mathcal{P}_{\omega,t}^{(\beta)}, \tag{23}$$

where

$$\mathcal{P}_{\omega,t}^{(\beta)} = \begin{cases} \mathcal{N}_{\omega,\sqcup}^{(\beta)} - \dagger_{\omega,\sqcup} \mathcal{M}_{\omega,\sqcup}^{(\beta-\infty)} & (\beta < 1) \\ \mathcal{M}_{\omega,t}^{(\beta)} - y_{\omega,t} \mathcal{M}_{\omega,t}^{(\beta-1)} & (1 \leq \beta \leq 2) \\ \mathcal{M}_{\omega,t}^{(\beta)} - y_{\omega,t} \mathcal{N}_{\omega,t}^{(\beta-1)} & (\beta > 2) \end{cases}, \tag{24}$$

$$\mathcal{M}_{\omega,t}^{(\beta)} = \frac{1}{\beta} \sum_k \alpha_{\omega,t,k} \eta_1 \left(\frac{f_{\omega,k} g_{k,t}}{\alpha_{\omega,t,k} \eta_1} \right)^{\beta} + \frac{1}{\beta} \sum_l \gamma_{\omega,t,l} \eta_2 \left(\frac{h_{\omega,l} u_{l,t}}{\gamma_{\omega,t,l} \eta_2} \right)^{\beta}, \tag{25}$$

$$\mathcal{N}_{\omega,t}^{(\beta)} = \sigma_{\omega,t}^{\beta-1} \left(z_{\omega,t} - \sigma_{\omega,t} \right) + \frac{\sigma_{\omega,t}^{\beta}}{\beta}. \tag{26}$$

The equality in (23) holds if and only if the auxiliary variables are set as follows:

$$\alpha_{\omega,t,k} = \frac{f_{\omega,k} g_{k,t}}{\sum_{k'} f_{\omega,k'} g_{k',t}}, \tag{27}$$

$$\gamma_{\omega,t,l} = \frac{h_{\omega,l} u_{l,t}}{\sum_{l'} h_{\omega,l'} u_{l',t}}, \tag{28}$$

$$\eta_1 = \frac{\sum_{k'} f_{\omega,k'} g_{k',t}}{\sum_{k'} f_{\omega,k'} g_{k',t} + \sum_{l'} h_{\omega,l'} u_{l',t}}, \tag{29}$$

$$\eta_2 = \frac{\sum_{l'} h_{\omega,l'} u_{l',t}}{\sum_{k'} f_{\omega,k'} g_{k',t} + \sum_{l'} h_{\omega,l'} u_{l',t}}, \tag{30}$$

$$\sigma_{\omega,t} = \sum\nolimits_{k'} f_{\omega,k'} g_{k',t} + \sum\nolimits_{l'} h_{\omega,l'} u_{l',t}. \tag{31}$$

Therefore, (27)–(31) are the update rules for auxiliary variables $\alpha_{\omega,t,k}$, $\gamma_{\omega,t,l}$, η_1, η_2, and $\sigma_{\omega,t}$, which correspond to (17).

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 (23)–(26), we can define the upper bound function $\mathcal{J}_2^+$ using auxiliary variables $\alpha_{\omega,t,k}$ and $\rho_{\omega,t}$ as

$$\mathcal{J}_2 \leq \mathcal{J}_2^+ = \sum_{\omega,t} \overline{i_{\omega,t}} \mathcal{S}_{\omega,t}^{(\beta_{\mathrm{R}})}, \tag{32}$$

where

$$\mathcal{S}_{\omega,t}^{(\beta_{\mathrm{R}})} = \begin{cases} \rho_{\omega,t}^{\beta_{\mathrm{R}}-1} \left(\sum_k f_{\omega,k} g_{k,t} - \rho_{\omega,t} \right) + \frac{\rho_{\omega,t}^{\beta_{\mathrm{R}}}}{\beta_{\mathrm{R}}} & (\beta_{\mathrm{R}} < 1) \\ \frac{1}{\beta_{\mathrm{R}}} \sum_k \alpha_{\omega,t,k} \left(\frac{f_{\omega,k} g_{k,t}}{\alpha_{\omega,t,k}} \right)^{\beta_{\mathrm{R}}} & (1 \leq \beta_{\mathrm{R}}) \end{cases}. \tag{33}$$

The equality in (32) holds if and only if the auxiliary variable $\alpha_{\omega,t,k}$ is set as (27) and $\rho_{\omega,t}$ is set as follows:

$$\rho_{\omega,t} = \sum_{k'} f_{\omega,k'} g_{k',t}. \tag{34}$$

Similarly to (27)–(31), (34) is the update rule for the auxiliary variable $\rho_{\omega,t}$.

Third, we define the upper-bound function for $\mathcal{J}_3$ using the auxiliary variable $\delta_{k,l,\omega} \geq 0$ that satisfies $\sum_\omega \delta_{k,l,\omega} = 1$ as

$$\mathcal{J}_3 \leq \mathcal{J}_3^+ = \sum_{k,l,\omega} \frac{f_{\omega,k}^2 h_{\omega,l}^2}{\delta_{k,l,\omega}}. \tag{35}$$

The equality in (35) holds if and only if the auxiliary variable is set as follows:

$$\delta_{k,l,\omega} = \frac{f_{\omega,k} h_{\omega,l}}{\sum_{\omega'} f_{\omega',k} h_{\omega',l}}. \tag{36}$$

Equation (36) is the update rule for the auxiliary variable $\delta_{k,l,\omega}$.

Finally, using (23), (32), and (35), we can define the upper-bound function $\mathcal{J}^+(\Theta, \hat{\Theta})$ as

$$\mathcal{J}(\Theta) \leq \mathcal{J}^+(\Theta, \hat{\Theta}) = \mathcal{J}_1^+ + \lambda \mathcal{J}_2^+ + \mu \mathcal{J}_3^+, \tag{37}$$

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

$$\sum_\omega i_{\omega,t} \left(\mathcal{V}_\beta - \mathcal{W}_\beta \right) + \lambda \mathcal{X}_{\beta_{\mathrm{R}}} = 0, \tag{38}$$

where

$$\mathcal{V}_\beta = \begin{cases} \sigma_{\omega,t}^{\beta-1} f_{\omega,k} & (\beta < 1) \\ g_{k,t}^{\beta-1} \left(\alpha_{k,\omega,t} \eta_1 \right)^{1-\beta} f_{\omega,k}^{\beta} & (1 \leq \beta) \end{cases}, \tag{39}$$

$$\mathcal{W}_\beta = \begin{cases} y_{\omega,t} g_{k,t}^{\beta-2} \left(\alpha_{k,\omega,t} \eta_1 \right)^{2-\beta} f_{\omega,k}^{\beta-1} & (\beta \leq 2) \\ y_{\omega,t} \sigma_{\omega,t}^{\beta-2} f_{\omega,k} & (2 < \beta) \end{cases}, \tag{40}$$

$$\mathcal{X}_{\beta_{\mathrm{R}}} = \begin{cases} \sum_\omega \overline{i_{\omega,t}} \rho_{\omega,t}^{\beta_{\mathrm{R}}-1} f_{\omega,k} & (\beta_{\mathrm{R}} < 1) \\ \sum_\omega \overline{i_{\omega,t}} f_{\omega,k} \left(\frac{f_{\omega,k} g_{k,t}}{\alpha_{\omega,t,k}} \right)^{\beta_{\mathrm{R}}-1} & (1 \leq \beta_{\mathrm{R}}) \end{cases}. \tag{41}$$

By solving (38) for $g_{k,t}$ assuming nonnegativity, we obtain

$$g_{k,t} = \begin{cases} \left(\frac{\sum_\omega i_{\omega,t} y_{\omega,t} \left(\alpha_{k,\omega,t} \eta_1 \right)^{2-\beta} f_{\omega,k}^{\beta-1}}{\sum_\omega i_{\omega,t} \sigma_{\omega,t}^{\beta-1} f_{\omega,k} + \lambda \mathcal{X}_{\beta_{\mathrm{R}}}} \right)^{\frac{1}{2-\beta}} & (\beta < 1) \\ \frac{\sum_\omega i_{\omega,t} y_{\omega,t} \left(\alpha_{k,\omega,t} \eta_1 \right)^{2-\beta} f_{\omega,k}^{\beta-1} g_{k,t}^{\beta-1}}{\sum_\omega i_{\omega,t} g_{k,t}^{\beta-1} \left(\alpha_{k,\omega,t} \eta_1 \right)^{1-\beta} f_{\omega,k}^{\beta} + \lambda \mathcal{X}_{\beta_{\mathrm{R}}}} & (1 \leq \beta \leq 2) \\ \left(\frac{\sum_\omega i_{\omega,t} y_{\omega,t} \sigma_{\omega,t}^{\beta-2} f_{\omega,k} g_{k,t}^{\beta-1}}{\sum_\omega i_{\omega,t} g_{k,t}^{\beta-1} \left(\alpha_{k,\omega,t} \eta_1 \right)^{1-\beta} f_{\omega,k}^{\beta} + \lambda \mathcal{X}_{\beta_{\mathrm{R}}}} \right)^{\frac{1}{\beta-1}} & (2 < \beta) \end{cases}. \tag{42}$$

This equation is one of the updates of the primary variables Θ and corresponds to (18). Then we can obtain more efficient update rules of $g_{k,t}$ by substituting the update rules of the auxiliary variables (27), (29), (31), and (34) into (42) as follows:

$$g_{k,t} \leftarrow g_{k,t} \left(\frac{\sum_\omega i_{\omega,t} y_{\omega,t} f_{\omega,k} z_{\omega,t}^{\beta-2}}{\sum_\omega i_{\omega,t} f_{\omega,k} z_{\omega,t}^{\beta-1} + \lambda R_G} \right)^{\varphi(\beta)}, \tag{43}$$

where R_G is given by

$$R_G = \sum_\omega \overline{i_{\omega,t}} f_{\omega,k} \left(\sum_{k'} f_{\omega,k'} g_{k',t} \right)^{\beta_{\mathrm{R}}-1}. \tag{44}$$

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

$$h_{\omega,l} \leftarrow h_{\omega,l}\left(\frac{\sum_t i_{\omega,t} y_{\omega,t} u_{l,t} z_{\omega,t}^{\beta-2}}{\sum_t i_{\omega,t} u_{l,t} z_{\omega,t}^{\beta-1} + 2\mu R_H}\right)^{\varphi(\beta)}, \quad (45)$$

$$u_{l,t} \leftarrow u_{l,t}\left(\frac{\sum_\omega i_{\omega,t} y_{\omega,t} h_{\omega,l} z_{\omega,t}^{\beta-2}}{\sum_\omega i_{\omega,t} h_{\omega,l} z_{\omega,t}^{\beta-1}}\right)^{\varphi(\beta)}, \quad (46)$$

where R_H is given by

$$R_H = \sum_k f_{\omega,k} \sum_{\omega'} f_{\omega',k} h_{\omega',l}. \quad (47)$$

The convergence of these update rules has been theoretically proven for all real values of β and β_{R} [16].

B. Experimental 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 has been discussed for signal separation [19], [20]. However, there has been no discussion on the optimal divergence for the extrapolation techniques using NMF. In this section, we experimentally analyze the extrapolation ability based on a statistical generation model of the observed data $\boldsymbol{Y}$, and determine the optimal divergence for basis extrapolation for various β and β_{R} values [21].

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 [22]. The minimization of $\mathcal{D}_\beta(y_{\omega,t}\|\vartheta)$ is equivalent to the maximization of $\exp(-\mathcal{D}_\beta(y_{\omega,t}\|\vartheta))$. Here, we can rewrite $\exp(-\mathcal{D}_\beta(y_{\omega,t}\|\vartheta))$ as

$$\exp\left(-\mathcal{D}_\beta\left(y_{\omega,t}\|\vartheta\right)\right) = \begin{cases} \frac{y_{\omega,t}}{\vartheta}\exp\left(-\frac{y_{\omega,t}}{\vartheta}+1\right) & (\beta=0) \\ \left(\frac{\vartheta e}{y_{\omega,t}}\right)^{y_{\omega,t}}\exp\left(-\vartheta\right) & (\beta=1) \\ \exp\left(-\frac{(y_{\omega,t}-\vartheta)^2}{2}\right) & (\beta=2) \\ \exp\left(\frac{\vartheta^{\beta-1}y_{\omega,t}}{\beta-1}-\frac{\vartheta^\beta}{\beta}\right) & (\beta\geq 3) \end{cases}, \quad (48)$$

where $\vartheta = \sum_k f_{\omega,k} g_{k,t}$ represents a parameter of the maximum likelihood estimation. A probability density function (p.d.f.) that corresponds to (48) is given by

$$y_{\omega,t} \sim p\left(y_{\omega,t}\right) = \begin{cases} \frac{1}{\vartheta_1}\exp\left(-\frac{y_{\omega,t}}{\vartheta_1}\right) & (\beta=0) \\ \frac{{\vartheta_2}^{y_{\omega,t}}}{\Gamma(y_{\omega,t}+1)}\exp\left(-\vartheta_2\right) & (\beta=1) \\ \frac{1}{\sqrt{2\pi\vartheta_3}}\exp\left(-\frac{(y_{\omega,t}-\vartheta_4)^2}{2{\vartheta_3}^2}\right) & (\beta=2) \\ C\exp\left(\frac{{\vartheta_5}^{\beta-1}y_{\omega,t}}{\beta-1}\right) & (\beta\geq 3) \end{cases}, \quad (49)$$

where $\Gamma(\cdot)$ is a gamma function. These generation models of $\beta = 0$, 1, and 2 are equivalent to exponential, Poisson, and Gaussian distributions, respectively. The generation models for $\beta \geq 3$ correspond to a distribution in which the probability increases exponentially with increasing $y_{\omega,t}$. Strictly, such a 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 C_{M} and define the corresponding p.d.f. with normalization coefficient C, which is given by

$$C = {\vartheta_5}^{\beta-1}(\beta-1)^{-1}\left(\exp\left(\frac{{\vartheta_5}^{\beta-1}}{\beta-1}C_{\mathrm{M}}\right)-1\right)^{-1}. \quad (50)$$

Using (49), we can generate the most probable spectrogram for each β.

2) Simulation Conditions: To analyze the net extrapolation ability, we simulated the spectrogram restoration task. In this simulation, we generated random i.i.d. values, which obey the corresponding generation model (49) 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 parameters of each p.d.f. to $\vartheta_1 = 1$, $\vartheta_2 = 5$, $\vartheta_3 = 10$, $\vartheta_4 = 50$, $\vartheta_5 = 2$, and $C_{\mathrm{M}} = 15$. These parameters were determined so as to generate 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 values without loss of generality. In addition, we used two types of data-missing pattern $\boldsymbol{I}$, in which 75% or 98% of the spectral grid points 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, $\boldsymbol{Y}_{\mathrm{target}} = \boldsymbol{Y}$ in (7) and (8). The number of supervised bases, K, was 100, which is the half the value of T, and the number of other bases, L, was 30. Therefore, the task was to reconstruct the 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 the sources-to-artifacts ratio (SAR) defined in [23] as the accuracy of the extrapolation. In this task, the observed signal $\boldsymbol{Y}$ does not have any interference sources. Therefore, SAR, which measures the absence of artificial distortion, is a good evaluation score for the restoration of the target signal. Here, the estimated signal $\hat{s}(t)$ is defined as

$$\hat{s}\left(t\right) = s_{\mathrm{target}}\left(t\right) + s_{\mathrm{interf}}\left(t\right) + s_{\mathrm{artif}}\left(t\right), \quad (51)$$

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 interference 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 formula for SAR is defined as

$$\mathrm{SAR} = 10\log_{10}\frac{\sum_t\left\{s_{\mathrm{target}}(t)+e_{\mathrm{interf}}(t)\right\}^2}{\sum_t {e_{\mathrm{artif}}(t)}^2}. \quad (52)$$

Fig. 6 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 (β_{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-

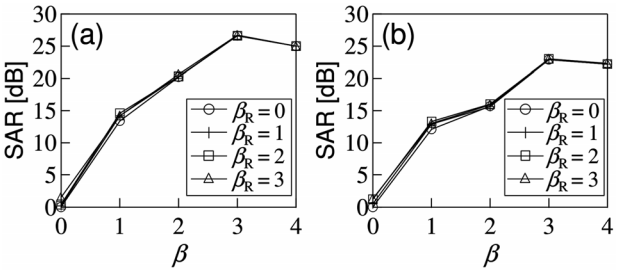

Fig. 6. Extrapolation abilities for (a) 75%-binary-masked data and (b) 98%-binary-masked data.

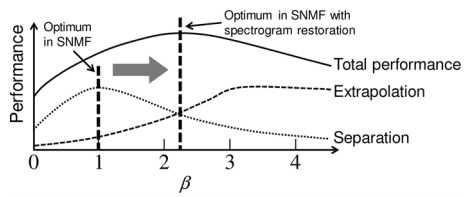

Fig. 7. Conceptual illustration of trade-off between separation and extrapolation abilities. The overall performance is highest when $\beta > 1$.

based music source separation because of the inherent sparsity of music spectrograms (e.g., $\beta = 1$ is recommended in [19], [20]). 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. 7. This issue will be confirmed experimentally in the next section.

C. Comparison Between Proposed Hybrid Method and Conventional Methods

1) Experimental Conditions: We conducted an objective evaluation to confirm the effectiveness of the proposed hybrid method described in the previous section. In this experiment, we compared the separation performance of six methods, namely, simple directional clustering [12], multichannel NMF [11] and its supervised version (supervised multichannel NMF), simple SNMF [19], naive hybrid method described in Section III-A1, and the proposed hybrid method including SNMF with spectrogram restoration after directional clustering, in terms of their ability to separate artificial and real-recorded music signals. The supervised multichannel NMF employs a priori training of the target spectral bases as well as SNMF and the hybrid methods, and we initialized the spatial covariance matrices of the supervised bases as the center direction for directional supervision of the target source. Also, we compared evaluation scores obtained with various β and β_{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 $\beta_{\mathrm{R}} = 0, 1, 2$. We used the same divergence (β) in the training and separation processes for the supervised methods.

In this evaluation, we conducted two experiments to consider artificial signal and real-recorded signal cases. We used stereo signals containing four melody parts (depicted in Fig. 8) with three compositions (C1–C3) of instruments as shown in Table I.

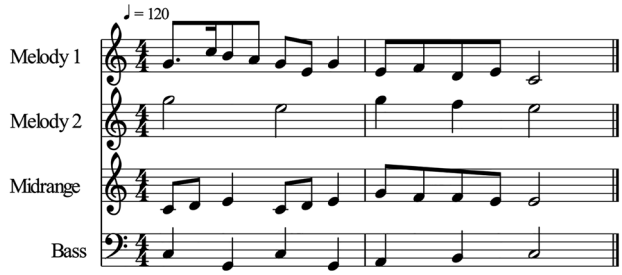

Fig. 8. Scores of each part.

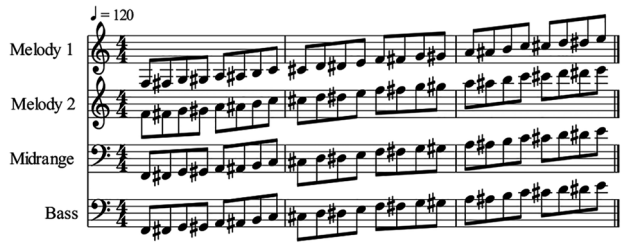

Fig. 9. Scores of each training sound that contain notes over two octaves. Note that only target instrumental sound is used in training stage.

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

The training signal $\boldsymbol{Y}_{\mathrm{target}}$ consisted of notes over two octaves that covered all the notes of the target instrument in the observed signal (see Fig. 9). This was artificially generated by a YAMAHA MU-1000 PCM-based MIDI synthesizer (hereafter referred to as *Tone Generator A*). Note that only the target instrument was trained in the training stage. We prepared three types of observed test signals $\boldsymbol{Y}$, namely, test signals generated by Tone Generator A, another type of PCM-based MIDI synthesizer Microsoft GS Wavetable Synth (hereafter referred to as *Tone Generator B*), and Garritan Personal Orchestra 4 (hereafter referred to as *Tone Generator C*). The test signal generated by Tone Generator A has the same timbre as the training sound, meaning that the best supervised bases were given for the separation task. The test signal generated by Tone Generator B provides different synthesized instrumental sounds, and that generated by Tone Generator C imitates more realistic sounds based on professionally recorded sample sounds. In addition, when using Tone Generator B and Tone Generator C, we added independent white Gaussian noises to the left and right channels of the observed signal $\boldsymbol{Y}$ with SNR $= 10$ dB to simulate background noise. In particular, these stereo signals were mixed down to a monaural format only when we evaluated 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 each in the left and right directions and two sources in the center direction based on the sine law (see Fig. 10(a)). The target instrument was always located in the center direction along with another interfering instrument,

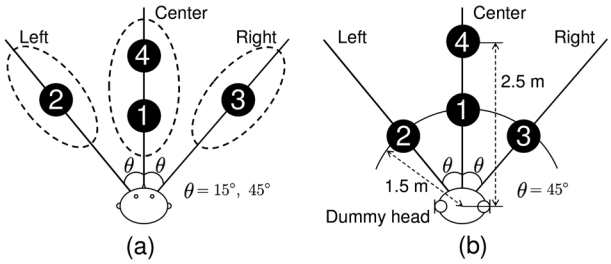

Fig. 10. Location of four sources with sine law used in (a) artificial signal and (b) real-recorded signal cases. Numbered black circles represent locations of instruments in stereo format. The angle of left- and right-side sources are $\theta = 15°, 45°$ in artificial signal case and with $\theta = 45°$ in real-recorded signal case.

and we prepared two patterns in which the left and right sources were located at $\theta = 15°$ and 45°, respectively. The sampling frequency of all the signals was 44.1 kHz. The spectrograms were computed using a 92-ms-long rectangular window with a 46 ms overlap shift. These STFT settings were determined so as to obtain sufficient frequency resolution. The number of iterations used for the training and separation were 500. The number of a priori bases, K, was set to 100 to prepare four bases for each of the training notes (24 notes). In addition, 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 parameter for the orthogonal penalty, μ, was set to 10000 because suitably chosen high value gives a good separation result [19]. The weighting parameter for the regularization term, λ, affects the extrapolation and quality of separated sound. In this experiment, λ was set to the optimal value based on the development dataset, which comprised the observed signals whose target was an oboe. The rest of the observed signals were used as a test dataset.

In the real-recorded signal case, we recorded each instrumental solo signal and the supervision sound, which were generated by Tone Generator A, using a NEUMANN KU 100 binaural microphone 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), respectively. The geometry of the loudspeaker and binaural microphone is shown in Fig. 10(b), where $\theta = 45°$. The target source and the supervision sound were always located at position No.1 in Fig. 10(b). The observed signal $\boldsymbol{Y}$ was produced by mixing these recorded signals at the same power. The other conditions were the same as those of the artificial signal case.

2) Experimental Results: We used the signal-to-distortion ratio (SDR) defined in [23] as the evaluation score. The formula for SDR is defined as

$$\mathrm{SDR} = 10\log_{10}\frac{\sum_t s_{\mathrm{target}}(t)^2}{\sum_t \{e_{\mathrm{interf}}(t) + e_{\mathrm{atif}}(t)\}^2}. \qquad (53)$$

SDR indicates the total evaluation score, which involves the quality of the separated target sound and the degree of separation.

Figs. 11–13 show the average SDR scores of the proposed hybrid method and the other methods for each divergence (β) and each regularization (β_{R}) in the artificial signal case with

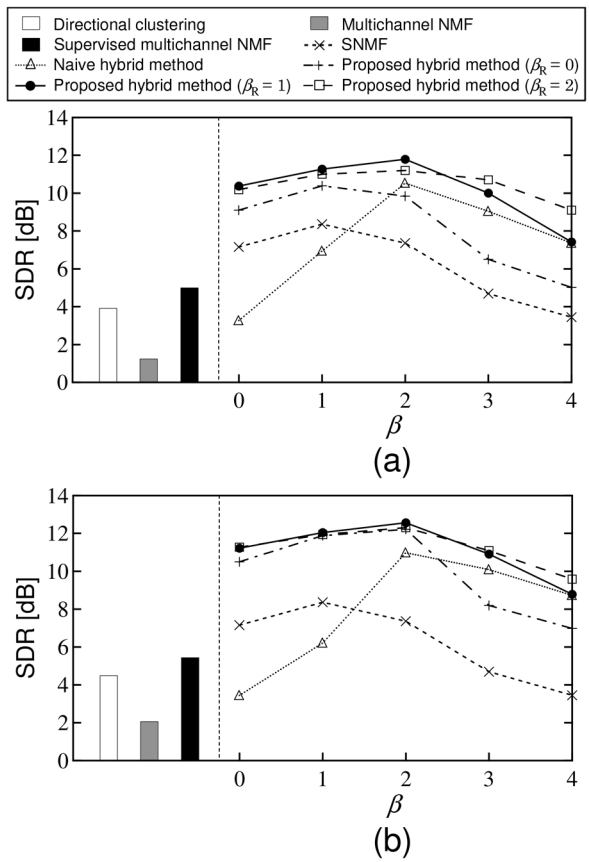

Fig. 11. Average SDR scores in artificial signal case using Tone Generator A when (a) $\theta = 15°$ and (b) $\theta = 45°$.

$\theta = 15°$ and $\theta = 45°$, where the four instruments are shuffled with 12 combinations in each of the compositions C1–C3, and three input signals whose target is the oboe are used as a development dataset. Therefore, these results are the averages of 33 input signal patterns (test dataset). Also, Fig. 14 shows the average SDR scores in the real-recorded signal case. From the SDR scores in Figs. 11—14, we can confirm that directional clustering and multichannel NMF do not have satisfactory performance because they cannot discriminate the sources in the same direction. Supervised multichannel NMF also cannot achieve satisfactory separation performance. For this reason, it is expected that (a) this method should be used to classify four source clusters with two-channel inputs, compared with two clusters (target and the rest; $\boldsymbol{FG}$ and $\boldsymbol{HU}$) in SNMF, and (b) as the number of clusters increases, this method should optimize more parameters such as spatial covariance matrices and latent variables, even if the target bases are given. In particular, the scores of multichannel NMF in Figs. 12–13 are markedly worse. In multichannel NMF, we must cluster the decomposed bases using their spatial covariance matrices to achieve the separation. However, if the diffuse noise exists, this method cannot separate the target signal well because such spatially uniform noise interferes with the clustering of decomposed bases. In contrast, SNMF-based methods can reduce such background noise by pushing them into the non-target component $\boldsymbol{HU}$ as interfering sources. This result shows an advantage of SNMF methods in terms of the robustness against the background noise. Also,

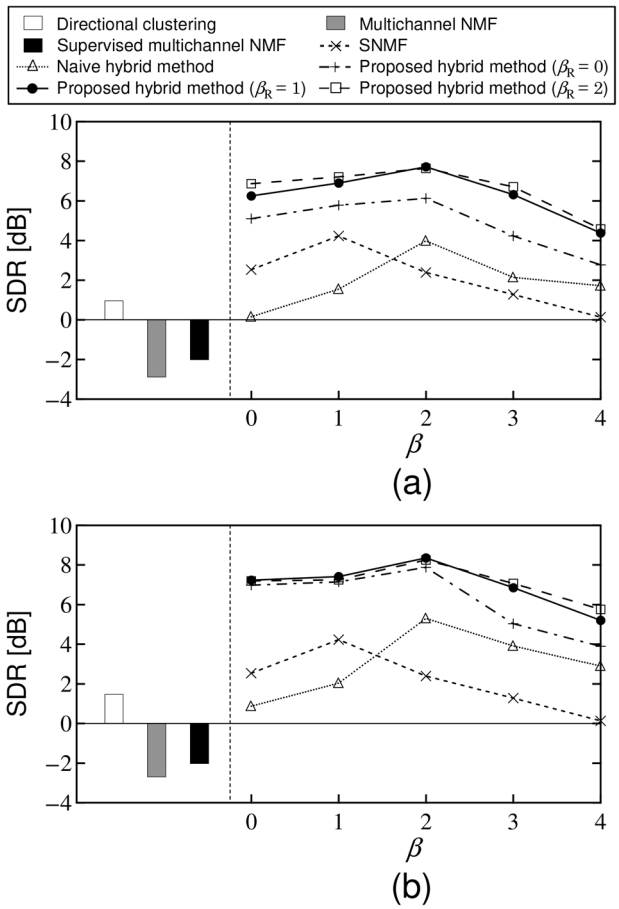

Fig. 12. Average SDR scores in artificial signal case using Tone Generator B with white noise when (a) $\theta = 15°$ and (b) $\theta = 45°$.

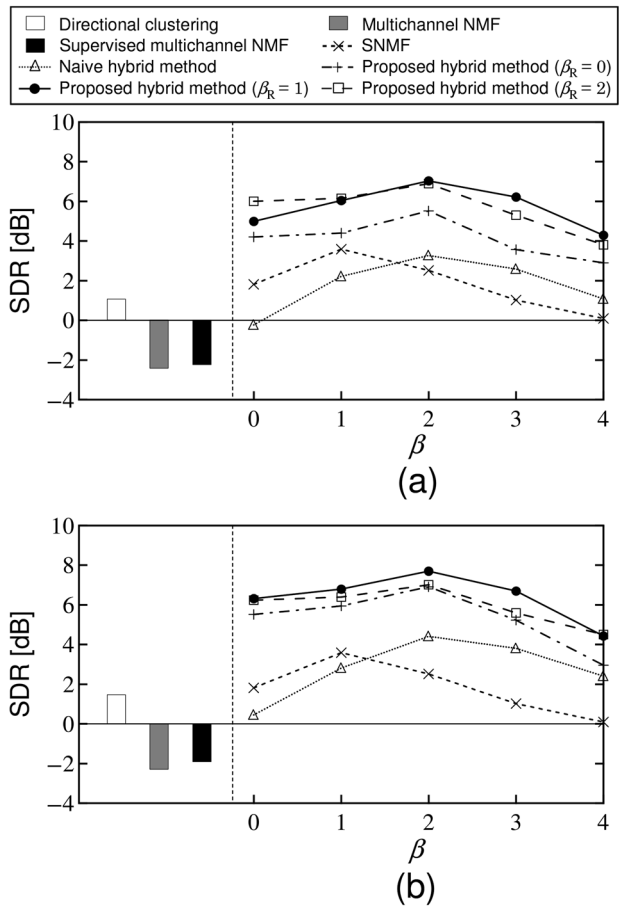

Fig. 13. Average SDR scores in artificial signal case using Tone Generator C with white noise when (a) $\theta = 15°$ and (b) $\theta = 45°$.

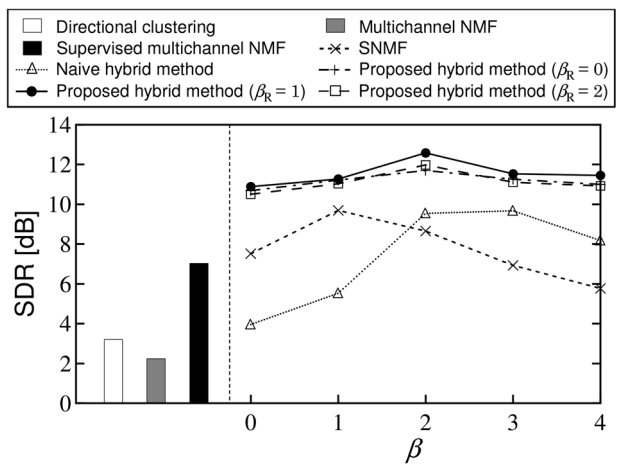

Fig. 14. Average SDR scores in real-recorded signal case when $\theta = 45°$.

in artificial signal cases, the sources are spatially arranged according to only the difference of amplitudes (sine law) between channels. Therefore, two sources at the center (Nos. 1 and 4 in Fig. 10(a)) have identical spatial properties. Thus, multichannel NMF cannot distinguish these sources and never achieves good separation in artificial signal cases. In contrast, for the real-recorded signals, supervised multichannel NMF achieves a certain level of separation because the two sources at the center direction have different room transfer functions (i.e., different spatial covariance matrices) as shown in Fig. 10(b).

The methods using SNMF 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 $\beta \leq 1$, where this hybrid method utilizes both directional clustering and SNMF. This is because the naive hybrid method is affected by spectral chasms and cannot reconstruct such lost components. Furthermore, we can confirm that the EUC-distance-based cost function ($\beta = 2$) is the optimal divergence for the proposed hybrid method, whereas KL-divergence ($\beta = 1$) is the best divergence even for conventional SNMF [19], [20]. This marked shift of the optimal divergence for SNMF with spectrogram restoration is due to the trade-off between the separation and extrapolation abilities, as predicted in Section III-B. In addition, the regularization with KL-divergence ($\beta_R = 1$) is slightly better than that with the other divergences but the difference is not significant, except when $\beta_R = 0$.

IV. SNMF with Spectrogram Restoration Based on Adaptive Divergence

A. Divergence Dependence on Local Chasm Condition

In Section III, we revealed the mechanism of the shift in the optimal divergence in the SNMF methods. This shift is due to

the trade-off between separation and extrapolation abilities. The optimal divergence for SNMF with spectrogram restoration depends on the density of spectral chasms in each time frame of the spectrogram obtained by the preceding directional clustering. Therefore, the optimal divergence temporally fluctuates because the spatial condition is not consistent in a general music signal, and the divergence of SNMF should be automatically changed in each time frame. To solve this problem, in this section, we propose a new scheme for framewise divergence selection to separate the target signal using the optimal divergence.

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

B. Cost Function and Update Rules

Considering the above-mentioned dependence of divergence on the local chasm condition, we propose to adapt the divergence in each frame of the spectrogram so that it is the optimal value according to the density of chasms in each frame r_t and a threshold value τ ($0 \leq \tau \leq 1$), where the density of chasms r_t can be calculated from the index matrix $\boldsymbol{I}$. A straightforward but naive extension to this purpose is to apply independent SNMF with spectrogram restoration to short-time-period data while 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 this problem and maintain sufficient dimensionality of the matrix, we propose a new batch SNMF with spectrogram restoration that includes an adaptive-divergence-based cost function covering the whole input matrix (see Fig. 15). The proposed cost function J_{m} is defined as

$$\mathcal{J}_{\mathrm{m}} = \sum_t \mathcal{J}_t, \tag{54}$$

$$J_t = \begin{cases} \sum_\omega i_{\omega,t} \mathcal{D}_{\beta=2}(y_{\omega,t} \| s^{(\mathrm{E})}_{\omega,t}) \\ +\lambda^{(\mathrm{E})} \sum_\omega \overline{i_{\omega,t}} \mathcal{D}_{\beta_{\mathrm{R}}}(0 \| \sum_k f^{(\mathrm{E})}_{\omega,k} g_{k,t}) \\ +\mu^{(\mathrm{E})} \|\boldsymbol{F}^{(\mathrm{E})\mathrm{T}} \boldsymbol{H}\|^2_{\mathrm{F}} (r_t \geq \tau) \\ \sum_\omega i_{\omega,t} \mathcal{D}_{\beta=1}(y_{\omega,t} \| s^{(\mathrm{K})}_{\omega,t}) \\ +\lambda^{(\mathrm{K})} \sum_\omega \overline{i_{\omega,t}} \mathcal{D}_{\beta_{\mathrm{R}}}(0 \| \sum_k f^{(\mathrm{K})}_{\omega,k} g_{k,t}) \\ +\mu^{(\mathrm{K})} \|\boldsymbol{F}^{(\mathrm{K})\mathrm{T}} \boldsymbol{H}\|^2_{\mathrm{F}} (r_t < \tau) \end{cases}, \tag{55}$$

$$s^{(*)}_{\omega,t} = \sum_k f^{(*)}_{\omega,k} g_{k,t} + \sum_n h_{\omega,n} u_{n,t}, \tag{56}$$

$$r_t = \frac{\sum_\omega \overline{i_{\omega,t}}}{\Omega}, \tag{57}$$

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 framewisely adaptive SNMF (hereafter referred to as *adaptive-divergence-based hybrid method*) to achieve both optimal separation and extrapolation.

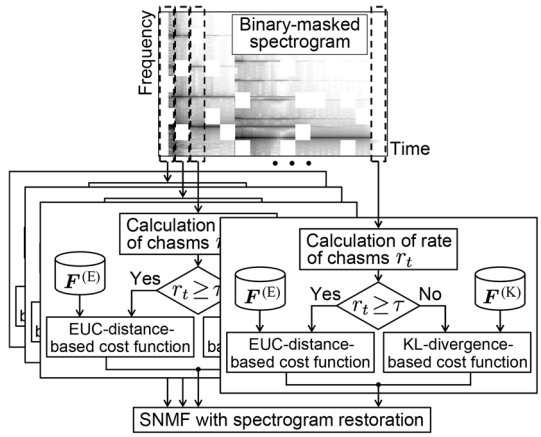

Fig. 15. Adaptive divergence algorithm in proposed method.

The update rules based on (54) are obtained by the auxiliary function approach. Similarly to in Section III-A2, we can design the upper-bound function $\mathcal{J}^+_{\mathrm{m}}$ using auxiliary variables $\zeta^{(*)}_{k,l,\omega} \geq 0$, $\kappa^{(*)}_{\omega,t,k} \geq 0$, $\gamma_{\omega,t,l} \geq 0$, $\varepsilon_1 \geq 0$, $\varepsilon_2 \geq 0$, and $\xi^{(*)}_{\omega,t} \geq 0$ that satisfy $\sum_\omega \zeta^{(*)}_{k,l,\omega} = 1$, $\sum_k \kappa^{(*)}_{\omega,t,k} = 1$, $\sum_l \gamma_{\omega,t,l} = 1$, and $\varepsilon_1 + \varepsilon_2 = 1$ as

$$\mathcal{J}_{\mathrm{m}} \leq \mathcal{J}^+_{\mathrm{m}} = \sum_t \mathcal{J}^+_t, \tag{58}$$

$$\mathcal{J}_t \leq \mathcal{J}^+_t = \begin{cases} \sum_\omega i_{\omega,t} \left(y^2_{\omega,t} + p_{\omega,t} + 2q_{\omega,t}\right) \\ +\lambda^{(\mathrm{E})} \sum_\omega \overline{i_{\omega,t}} \mathcal{R}^{(\mathrm{E})}_{\beta_{\mathrm{R}}} \\ +\mu^{(\mathrm{E})} \sum_{k,l,\omega} \frac{f^{(\mathrm{E})2}_{\omega,k} h^2_{\omega,l}}{\zeta^{(\mathrm{E})}_{k,l,\omega}} (r_t \geq \tau) \\ \sum_\omega i_{\omega,t} \left(-y_{\omega,t} \sum_{k,l} \kappa^{(\mathrm{K})}_{\omega,t,k} \gamma_{\omega,t,l} \mathcal{Q} + c\right) \\ +\lambda^{(\mathrm{K})} \sum_\omega \overline{i_{\omega,t}} \mathcal{R}^{(\mathrm{K})}_{\beta_{\mathrm{R}}} \\ +\mu^{(\mathrm{K})} \sum_{k,l,\omega} \frac{f^{(\mathrm{K})2}_{\omega,k} h^2_{\omega,l}}{\zeta^{(\mathrm{K})}_{k,l,\omega}} (r_t < \tau) \end{cases} \tag{59}$$

where

$$p_{\omega,t} = \sum_k \frac{f^{(\mathrm{E})2}_{\omega,k} g^2_{k,t}}{\kappa^{(\mathrm{E})}_{\omega,t,k}} + \sum_l \frac{h_{\omega,l} u_{l,t}}{\gamma_{\omega,t,l}}, \tag{60}$$

$$q_{\omega,t} = \left(\sum_k f^{(\mathrm{E})}_{\omega,k} g_{k,t}\right) \left(\sum_l h_{\omega,l} u_{l,t}\right) - y_{\omega,t} \sum_k f^{(\mathrm{E})}_{\omega,k} g_{k,t} - y_{\omega,t} \sum_l h_{\omega,l} u_{l,t}, \tag{61}$$

$$\mathcal{R}^{(*)}_{\beta_{\mathrm{R}}} = \begin{cases} \xi^{(*)\beta_{\mathrm{R}}-1}_{\omega,t} \left(\sum_k f^{(*)}_{\omega,k} g_{k,t} - \xi^{(*)}_{\omega,t}\right) + \frac{\xi^{(*)\beta_{\mathrm{R}}}_{\omega,t}}{\beta_{\mathrm{R}}} & (\beta_{\mathrm{R}} < 1) \\ \frac{1}{\beta_{\mathrm{R}}} \sum_k \kappa^{(*)}_{\omega,t,k} \left(\frac{f^{(*)}_{\omega,k} g_{k,t}}{\kappa^{(*)}_{\omega,t,k}}\right)^{\beta_{\mathrm{R}}} & (1 \leq \beta_{\mathrm{R}}) \end{cases}, \tag{62}$$

$$\mathcal{Q} = \varepsilon_1 \log \Phi + \varepsilon_2 \log \Psi, \tag{63}$$

$$c = -y_{\omega,t} \sum_{k,l} \kappa^{(\mathrm{K})}_{\omega,t,k} \gamma_{\omega,t,l} \cdot \left(\log \kappa^{(\mathrm{K})}_{\omega,t,k} \gamma_{\omega,t,l} + \varepsilon_1 \log \varepsilon_1 + \varepsilon_2 \log \varepsilon_2 \right), \tag{64}$$

$$\Phi = \gamma_{\omega,t,l} f^{(\mathrm{K})}_{\omega,k} g_{k,t}, \tag{65}$$

$$\Psi = \kappa^{(\mathrm{K})}_{\omega,t,k} h_{\omega,l} u_{l,t}. \tag{66}$$

The equality in (59) holds if and only if the auxiliary variables are set as in (28) and as follows:

$$\zeta^{(*)}_{k,l,\omega} = \frac{f^{(*)}_{\omega,k} h_{\omega,l}}{\sum_{\omega'} f^{(*)}_{\omega',k} h_{\omega',l}}, \tag{67}$$

$$\kappa^{(*)}_{\omega,t,k} = \frac{f^{(*)}_{\omega,k} g_{k,t}}{\sum_{k'} f^{(*)}_{\omega,k'} g_{k',t}}, \tag{68}$$

$$\varepsilon_1 = \frac{\Phi}{\Phi + \Psi}, \tag{69}$$

$$\varepsilon_2 = \frac{\Psi}{\Phi + \Psi}, \tag{70}$$

$$\xi^{(*)}_{\omega,t} = \sum_k f^{(*)}_{\omega,k} g_{k,t}. \tag{71}$$

The update rules are obtained as follows by differentiating the upper-bound function (58) w.r.t. each objective variable and substituting of the equality conditions (67)–(71);

$$g_{k,t} \leftarrow \begin{cases} g_{k,t} \cdot \dfrac{\sum_\omega i_{\omega,t} y_{\omega,t} f^{(\mathrm{E})}_{\omega,k}}{\sum_\omega i_{\omega,t} f^{(\mathrm{E})}_{\omega,k} s^{(\mathrm{E})}_{\omega,t} + \lambda^{(\mathrm{E})} R^{(\mathrm{E})}_G} & (r_t \geq \tau) \\ g_{k,t} \cdot \dfrac{\sum_\omega i_{\omega,t} y_{\omega,t} f^{(\mathrm{K})}_{\omega,k} s^{(\mathrm{K})-1}_{\omega,t}}{\sum_\omega i_{\omega,t} f^{(\mathrm{K})}_{\omega,k} + \lambda^{(\mathrm{K})} R^{(\mathrm{K})}_G} & (r_t < \tau) \end{cases}, \tag{72}$$

$$h_{\omega,l} \leftarrow h_{\omega,l} \cdot \frac{\sum_t i_{\omega,t} y_{\omega,t} u_{l,t} N_{\omega,t}}{\sum_t i_{\omega,t} u_{l,t} D_{\omega,t} + P_{\omega,l}}, \tag{73}$$

$$u_{l,t} \leftarrow \begin{cases} u_{l,t} \cdot \dfrac{\sum_\omega i_{\omega,t} y_{\omega,t} h_{\omega,l}}{\sum_\omega i_{\omega,t} h_{\omega,l} s^{(\mathrm{E})}_{\omega,t}} & (r_t \geq \tau) \\ u_{l,t} \cdot \dfrac{\sum_\omega i_{\omega,t} y_{\omega,t} h_{\omega,l} s^{(\mathrm{E})-1}_{\omega,t}}{\sum_\omega i_{\omega,t} h_{\omega,l}} & (r_t < \tau) \end{cases}, \tag{74}$$

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

$$R^{(*)}_G = \sum_\omega \overline{i_{\omega,t}} f^{(*)}_{\omega,k} \left(\sum_{k'} f^{(*)}_{\omega,k'} g_{k',t} \right)^{\beta_{\mathrm{R}}-1}, \tag{75}$$

$$N_{\omega,t} = \begin{cases} 1 & (r_t \geq \tau) \\ s^{(\mathrm{K})-1}_{\omega,t} & (r_t < \tau) \end{cases}, \tag{76}$$

$$D_{\omega,t} = \begin{cases} s^{(\mathrm{E})}_{\omega,t} & (r_t \geq \tau) \\ 1 & (r_t < \tau) \end{cases}, \tag{77}$$

$$P_{\omega,l} = \begin{cases} \mu^{(\mathrm{E})} \sum_k f^{(\mathrm{E})}_{\omega,k} \sum_{\omega'} f^{(\mathrm{E})}_{\omega',k} h_{\omega',l} & (r_t \geq \tau) \\ \mu^{(\mathrm{K})} \sum_k f^{(\mathrm{K})}_{\omega,k} \sum_{\omega'} f^{(\mathrm{K})}_{\omega',k} h_{\omega',l} & (r_t < \tau) \end{cases}. \tag{78}$$

The update rules of SNMF with spectrogram restoration based on adaptive divergence are defined as (72)–(74).

C. Evaluation Experiment

1) Experimental Conditions: To confirm the effectiveness of the proposed algorithm, we compared seven methods, namely, SNMF based on KL-divergence and EUC-distance [19], simple directional clustering [12], multichannel NMF [11] and its supervised version, the conventional hybrid method based on KL-divergence and EUC-distance, the online hybrid method described in Section IV-B, and the proposed hybrid method that uses adaptive divergence.

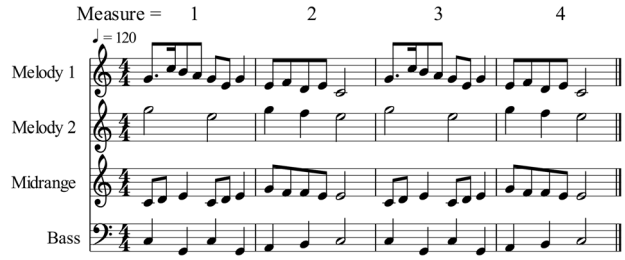

Fig. 16. Scores of each part. The observed signal consists of four measures.

TABLE II
SPATIAL CONDITIONS OF EACH DATASET

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

In this experiment, similarly to in Section III-C1, we produced artificial and real-recorded stereo signals containing four melody parts (depicted in Fig. 16) with the three compositions (C1–C3) of instruments shown in Table I. The artificial training and observed signals were generated with the same conditions in Section III-C1. These stereo signals were mixed down to a monaural format only when we evaluated 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. Note that the target signal was always located in the center direction along with another interference signal as shown in Fig. 10, and the left- and right-side interference signals were instantaneously moved to the center direction in the middle of the song for SP1–SP3. In the hybrid method, a large number of chasms were produced by directional clustering in the measures with $\theta = 45°$ compared with those with $\theta = 0°$. Therefore, we expected that EUC-distance-based hybrid method would be suitable for the dataset of SP4 rather than the dataset of SP1. The threshold value τ was set to 20%, which appears to be relatively small. This is because the separated sound quality is particularly important in music signal separation and the spectral chasms should be actively extrapolated. The type of regularization was $\beta_{\mathrm{R}} = 1$. The other experimental conditions were the same as those in Section III-C1.

2) Experimental Results: Fig. 17 shows the average SDR scores for each method and each dataset pattern. These results are the averages of 33 input signal patterns, similarly to in Section III-C1. The SDR scores of SNMF are the same for all datasets because the input signals for SNMF were mixed down to a monaural format.

From these results, the KL-divergence-based hybrid method achieves high separation accuracy for the datasets 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.

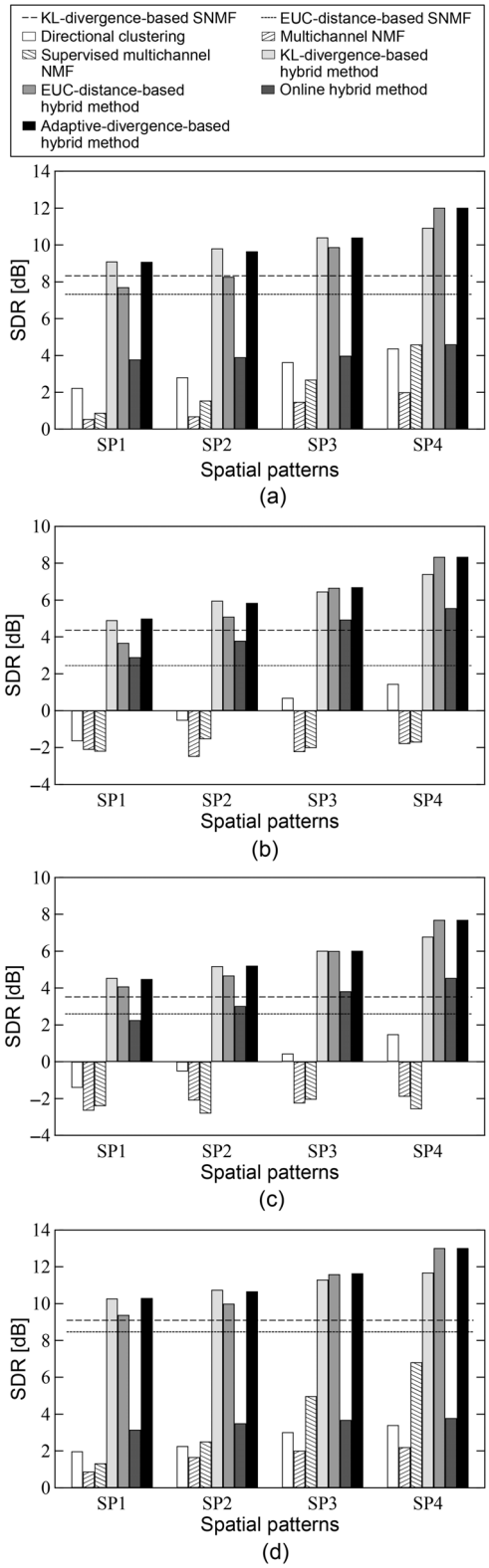

Fig. 17. Average SDR scores of each method and each spatial condition in (a) artificial signal case using Tone Generator A, (b) artificial signal case using Tone Generator B with white noise, (c) artificial signal case using Tone Generator C with white noise, and (d) real-recorded signal case.

This dataset has many spectral chasms because the signals are mixed with a wide panning angle ($\theta = 45°$), which yields many chasms, and high extrapolation ability is required. In addition, the proposed hybrid method with adaptive divergence achieves better separation for all datasets 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.

D. Experimental Comparison Between Adaptive-Divergence-Based Hybrid Method and Another Strategy

1) Parallel Divergence Method: As another means of applying multiple divergence to a whole spectrogram (batch method), the following method can also be considered for the adaptation of divergence. First, we divide the whole spectrogram $\boldsymbol{Y}$ into two parts, $\boldsymbol{Y}^{(\mathrm{E})}$ and $\boldsymbol{Y}^{(\mathrm{K})}$, based on the density of chasms r_t and threshold τ, where $\boldsymbol{Y}^{(\mathrm{E})}$ consists of the frames with r_t greater than τ and $\boldsymbol{Y}^{(\mathrm{K})}$ consists of the other frames. Then, we apply the EUC-distance-based and KL-divergence-based proposed methods to $\boldsymbol{Y}^{(\mathrm{E})}$ and $\boldsymbol{Y}^{(\mathrm{K})}$, respectively, in parallel. Finally, the separated whole spectrogram is reconstructed by concatenating the frames of the separated spectrograms in the original order. Hereafter we refer to this method as *parallel-divergence-based hybrid method*. In this method, the update rules of $g_{k,t}$ and $u_{l,t}$ are equivalent to (72) and (74), respectively. The difference between the parallel- and adaptive-divergence-based hybrid methods is how to deal with the interference matrix $\boldsymbol{HU}$. In the adaptive-divergence-based hybrid method, a single $\boldsymbol{HU}$ is optimized over all the frames in $\boldsymbol{Y}$ (thus, the dimensionality of $\boldsymbol{HU}$ is identical to that of $\boldsymbol{Y}$). On the other hand, the parallel-divergence-based hybrid method prepares two interference matrices, $\boldsymbol{H}^{(\mathrm{E})}\boldsymbol{U}^{(\mathrm{E})}$ and $\boldsymbol{H}^{(\mathrm{K})}\boldsymbol{U}^{(\mathrm{K})}$ for $\boldsymbol{Y}^{(\mathrm{E})}$ and $\boldsymbol{Y}^{(\mathrm{K})}$, respectively, whose frames are disjoint and whose dimensionality is reduced compared with that of $\boldsymbol{Y}$. Generally speaking, in the SNMF-based methods, the dimensionality of the input spectrogram affects the separation accuracy. This is because the interference matrix $\boldsymbol{HU}$ becomes an effective low-rank representation that ensures the success of separation when the number of frames increases. Therefore, we expect the parallel-divergence-based hybrid method to underperform compared with the adaptive-divergence-based hybrid method because the numbers of frames in $\boldsymbol{Y}^{(\mathrm{E})}$ and $\boldsymbol{Y}^{(\mathrm{K})}$ are small. This phenomenon is also expected to become more apparent as the number of frames in $\boldsymbol{Y}$ decreases. This will be experimentally confirmed in the next subsection.

2) Conditions and Results: We compared the two proposed hybrid methods based on adaptive and parallel divergence. We used three different lengths of the observed signals, which consist of two, three, and four measures with four melody parts (depicted in Fig. 16). As the spatial conditions, SP1 and SP2 were generated for the two-measure signal, SP1–SP3 were generated for the three-measure signal, and SP1–SP4 were generated for the four-measure signal, which were used in addition to the spatial conditions in Section IV-C. Similarly to in Section IV-C1, we produced artificial and real-recorded stereo signals containing the three compositions (C1–C3) of instruments shown in Table I. For the parallel divergence method, we set the number of bases in $\boldsymbol{H}^{(\mathrm{E})}$ and $\boldsymbol{H}^{(\mathrm{K})}$ to 30. The other experimental conditions were the same as those in Section IV-C1.

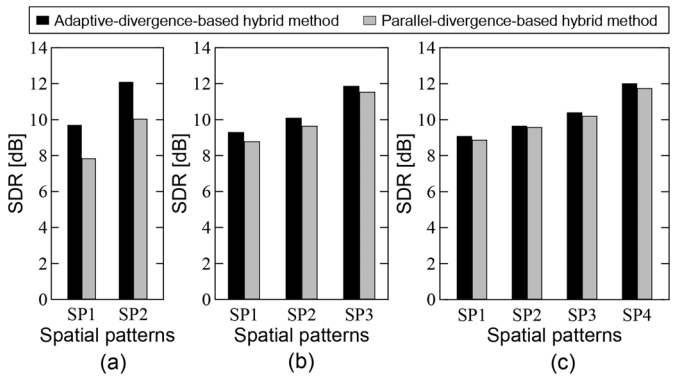

Fig. 18. Average SDR scores of each method and each spatial condition in artificial signal case using Tone Generator A: (a) two-measure-signal, (b) three-measure-signal, and (c) four-measure-signal.

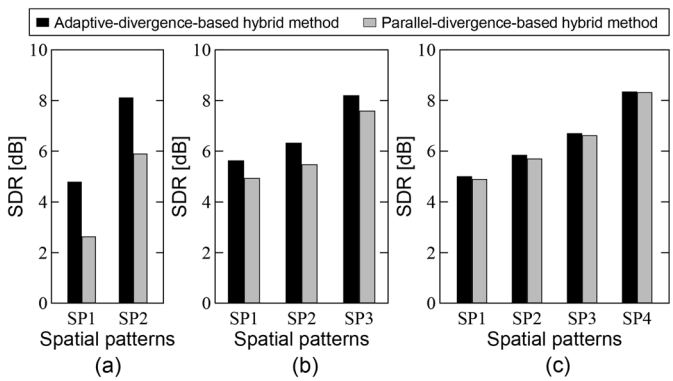

Fig. 19. Average SDR scores of each method and each spatial condition in artificial signal case using Tone Generator B with white noise: (a) two-measure-signal, (b) three-measure-signal, and (c) four-measure-signal.

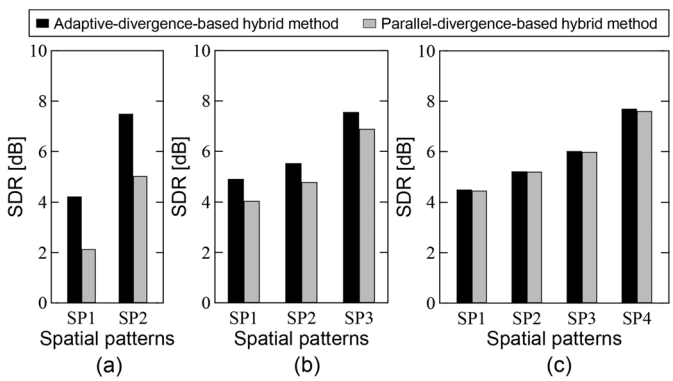

Fig. 20. Average SDR scores of each method and each spatial condition in artificial signal case using Tone Generator C with white noise: (a) two-measure-signal, (b) three-measure-signal, and (c) four-measure-signal.

Figs. 18–21 show the average SDR scores for each method and each case. From these results, we can confirm that the adaptive-divergence-based hybrid method outperforms the parallel-divergence-based hybrid method for all tasks. The degree of superiority is marked when the input signal is short (e.g., compare Fig. 18(a) with Fig. 18(c)), as predicted in the previous subsection.

V. Conclusion

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

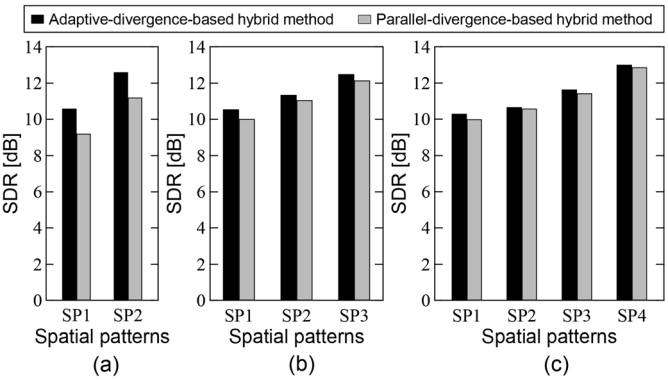

Fig. 21. Average SDR scores of each method and each spatial condition in real-recorded signal case: (a) two-measure-signal, (b) three-measure-signal, and (c) four-measure-signal.

Secondly, from experimental analysis based on the generation model of the signal, it was revealed that the optimal divergence in SNMF with spectrogram restoration, which attempts to fit the trained bases using spectral components except for the chasms, is shifted to an anti-sparse divergence rather than KL-divergence. This was due to the fact that a trade-off exists between the separation and extrapolation abilities in SNMF. An experiment evaluating 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 adaptive divergence. The proposed method adapts the divergence in each frame to the optimal one using a threshold value for the density of chasms to separate and extrapolate the target signal with high accuracy. Experimental results showed that our proposed method can achieve high separation accuracy under all spatial conditions.

Appendix A
Hybrid Method with Soft Directional Mask

Instead of employing a hard clustering method for directional separation, we can apply soft directional clustering, which provides a soft time-frequency mask. Various types of soft mask have been proposed [24]–[27]. Here, we generated a soft mask using MESSL as proposed in [27]. In MESSL, we chose the specific set of parameters, Θ_{11}, with frequency-independent Gaussian model for interaural level and phase differences [27], which gave the best SDR score in our task.

To evaluate the efficacy of the soft-mask-based hybrid method, we conducted the experiment described in Section III-C with a real-recorded signal case. Simple soft directional clustering [27] gives an SDR of less than 5 dB. Fig. 22 shows a comparison of the SDR between the proposed hybrid methods using hard and soft directional masks with SNMF. From these results, the hybrid method with a soft mask slightly improves the SDR, although a clear improvement is

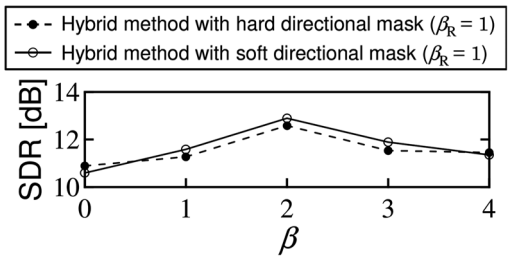

Fig. 22. Average SDR scores of hybrid methods with hard and soft masks in real-recorded signal case.

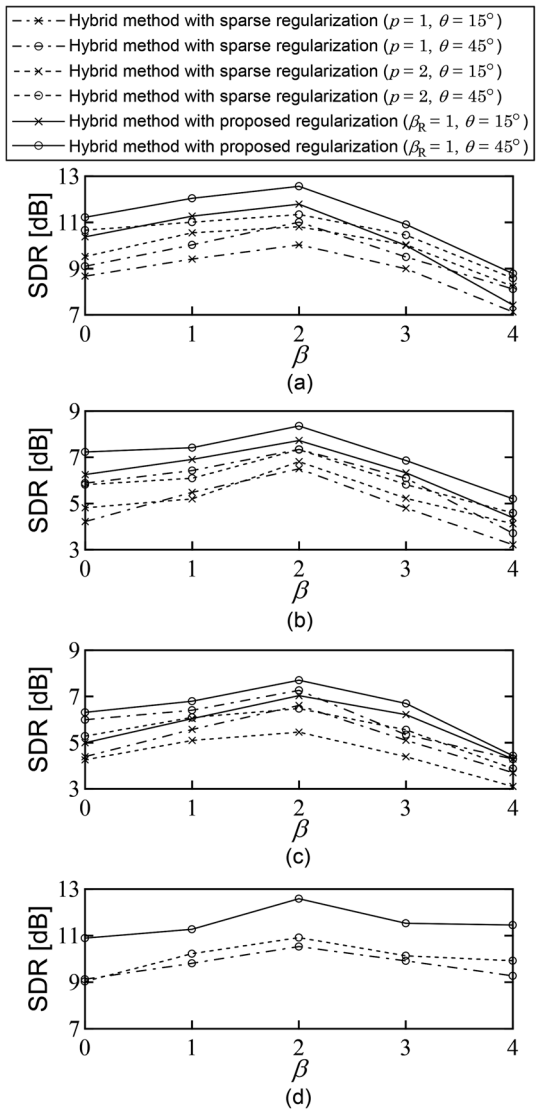

Fig. 23. Average SDR scores with sparse and proposed regularizations in (a) artificial signal case using Tone Generator A, (b) artificial signal case using Tone Generator B with white noise, (c) artificial signal case using Tone Generator C with white noise, and (d) real-recorded signal case.

not obtained. This is because even if all the target components are lost in some STFT grid points as a result of binary masking, SNMF with spectrogram restoration can reconstruct the lost target components by basis extrapolation. Thus, the restoration ability does not depend on the type of directional mask.

Appendix B
Sparse Regularization for Activation Matrix

To avoid the extrapolation error, the sparse regularization for $\boldsymbol{G}$ can also be used. This can be achieved by substituting the following equation for $\mathcal{J}_2$ term in (19):

$$\mathcal{J}_2 = \sum_{k,t} g_{k,t}^p, \tag{79}$$

where $p = 1$ or 2, which corresponds to L_1 or L_2 norm of $\boldsymbol{G}$. This penalty term increases sparseness of $\boldsymbol{G}$.

To compare the separation performance in the cases of (79) and (21), we conducted the same experiment in Section III-C. The results are shown in Fig. 23. From these results, sparse regularization for $\boldsymbol{G}$ can also avoid the extrapolation error but the scores do not outperform that of the proposed penalty. This is because the proposed penalty only affects the frames that have many chasms whereas the penalty in (79) imposes unnecessary sparseness to all the frames.

References

[1] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. 16th Eur. Signal Process. Conf.*, 2008.

[2] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proc. 8th Int. Conf. Music Inf. Retrieval*, 2007, pp. 375–378.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, vol. 13, pp. 556–562.

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

[5] A. Ozerov, C. Fevotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2009, pp. 121–124.

[6] 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," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 5365–5368.

[7] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. 7th Int. Conf. Ind. Compon. Anal. Signal Separat.*, 2007, pp. 414–421.

[8] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[9] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using NMF," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[10] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[11] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 261–264.

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

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

[14] P. Smaragdis and B. Raj, "Example-driven bandwidth expansion," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2007, pp. 135–138.

[15] S. Eguchi and Y. Kano, Robustifying maximum likelihood estimation Inst. of Statist. Math., Tech. Rep., 2001.

[16] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with beta-divergence," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2010, pp. 283–288.
[17] D. Kitamura, H. Saruwatari, Y. Iwao, K. Shikano, K. Kondo, and Y. Takahashi, "Superresolution-based stereo signal separation via supervised nonnegative matrix factorization," in *Proc. Digital Signal Process.*, 2013, pp. T3C–2.
[18] D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, "Superresolution-based stereo signal separation with regularization of supervised basis extrapolation," in *Proc. 3D Systems Applicat.*, 2013, pp. S10–4.
[19] 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. Fundam. Electron., Commun. Comput. Sci.*, vol. E97-A, no. 5, pp. 1113–1118, 2014.
[20] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Proc. Irish Signals Syst. Conf.*, 2009.
[21] D. Kitamura, H. Saruwatari, S. Nakamura, Y. Takahashi, K. Kondo, and K. Kameoka, "Divergence optimization in nonnegative matrix factorization with spectrogram restoration for multichannel signal separation," in *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, 2014, pp. 92–96.
[22] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, vol. 2009, pp. 1–17, 2009.
[23] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[24] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Trans. Speech Commun.*, vol. 48, pp. 1486–1501, 2006.
[25] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 58–67, Jan. 2006.
[26] N. Madhu and J. Wouters, "Localisation-based, situation-adaptive mask generation for source separation," in *Proc. 4th Int. Symp. Commun., Control, Signal Process.*, 2010.
[27] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

Daichi Kitamura (M'15) graduated from Nara institute of Science and Technology, Japan, in 2014 and received the M.E. degree. He is currently pursuing the Ph.D. degree at The Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan. His research interests include source separation and music signal processing. Mr. Kitamura is a member of the IEICE and the Acoustical Society of Japan.

Hiroshi Saruwatari (M'00) received the B.E., M.E., and Ph.D. degrees from Nagoya University, Japan, in 1991, 1993, and 2000, respectively. After joining Nara Institute of Science and Technology, he is currently a Professor of The University of Tokyo, Japan. His research interests include speech and acoustic signal processing. He is a member of the IEICE, Japan VR Society, and the Acoustical Society of Japan.

Hirokazu Kameoka received B.E., M.S., and Ph.D. degrees all from The University of Tokyo, Japan, in 2002, 2004, and 2007, respectively. He is currently a Research Scientist at NTT CS Laboratories and a Visiting Associate Professor at The University of Tokyo. His research interests include speech and music processing, and machine learning. He is the Information Processing Society of Japan, and the Acoustical Society of Japan.

Yu Takahashi (S'07) received the B.E. degree in information engineering from the Himeji Institute of Technology, Japan, in 2005 and the M.E. and Ph.D. degrees in information science from the Nara Institute of Science and Technology, Japan, in 2007 and 2010, respectively. He joined Yamaha Co., Ltd. in 2010. His research interests include array signal processing and blind source separation. He is a member of the Acoustical Society Japan and a member of the Japanese Society for Artificial Intelligence.

Kazunobu Kondo received the B.E., M.E. and Ph.D. degrees from Nagoya University, Japan, in 1991, 1993, and 2014, respectively. He joined the Electronics Development Center, Yamaha Co., Ltd. in 1993. He is currently an Assistant Manager of Yamaha Research and Development Division. His research interests include blind source separation and noise reduction. He is a member of the IEICE and the Acoustical Society of Japan.

Satoshi Nakamura received the B.S. from Kyoto Institute of Technology in 1981 and the Ph.D. from Kyoto University in 1992. He is currently a Professor of Nara Institute of Science and Technology, Japan. His research interests include speech and language processing. He has been Elected Board Member of ISCA and IEEE Signal Processing Magazine Editorial Board Member. He is a member of the Acoustical Society of Japan.