Vol.2013-MUS-99 No.14
2013/5/11

Importance of Regularization in Superresolution-Based Multichannel Signal Separation with Nonnegative Matrix Factorization

Daichi Kitamura[1,a)] Hiroshi Saruwatari[1] Kiyohiro Shikano[1] Kazunobu Kondo[2]
Yu Takahashi[2]

Abstract: In this paper, we address a multichannel signal separation problem and propose a new hybrid method utilizing both directional clustering and supervised nonnegative matrix factorization (NMF) with norm regularization term. Conventional hybrid method concatenating supervised NMF after directional clustering has a problem that the extracted signal suffers from considerable spectral distortion because directional clustering yields spectral chasms. To solve this problem, we propose a new supervised NMF algorithm that regards the spectral chasms as unseen observations and reconstructs the target source components via spectrogram extrapolation. In addition, this paper addresses an issue on importance of regularization introduced in the superresolution procedure. Our experimental results show that the proposed hybrid method with regularization greatly improves the separation performance for stereo signals.

1. Introduction

In recent years, music signal separation based on nonnegative matrix factorization (NMF) [1], which is a type of sparse representation algorithm, has been a very active area of signal processing research [2], [3]. NMF for acoustical signals decomposes an input spectrogram into the product of a spectral basis matrix and its activation matrix. In particular, supervised NMF (SNMF) [4], [5], which includes a priori training with some sample sounds of a target instrument, can extract the target signal to some extent, particularly in the case of a small number of instruments. However, for the case of a mixture consisting of many sources, such as more realistic musical tunes, the source extraction performance is markedly degraded when only single-channel observation is available.

Multichannel NMF, which is a natural extension of NMF for a stereo or multichannel music signal, has been proposed as an unsupervised method [6], [7]. 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.

As another means for addressing multichannel signal separation, directional clustering has also been proposed as an unsupervised method [8], [9]. This method quantizes directional information via time-frequency binary masking under the assumption that the sources are completely sparse in the time-frequency domain. However, there is an inherent problem that sources located in the same direction cannot be separated using the directional information. To cope with this problem, a hybrid method for multichannel signal separation, which concatenates SNMF after directional clustering, has been proposed [10]. However, this hybrid method also has a problem that the extracted signal suffers from considerable distortion because the signal obtained by directional clustering has many spectral chasms. This results in the cascaded SNMF being forced to incorrectly mimic such artificial spectral chasms.

In this paper, we propose a new SNMF algorithm for the hybrid method. Using index information generated by binary masking, the proposed SNMF regards the spectral chasms as *unseen* observations, and finally reconstructs the target source components via spectrogram extrapolation using supervised bases. In other words, the proposed method can be categorized as *superresolution* because the degraded resolution in the time-frequency domain resulting from the preceding binary masking can be recovered. In addition, this paper addresses an issue on importance of *regularization* introduced in the superresolution procedure. Our experimental results show that the proposed method outperforms several conventional methods and that the distortion of the extracted signal can be mitigated by the effectiveness of the superresolution-based method.

2. Conventional Method

2.1 Overview of Penalized SNMF

The unsupervised NMF approaches have difficulty in clustering decomposed spectral patterns into a specific target instrumental sound. Furthermore, each basis may be forced to include a multi-instrumental spectral pattern. To solve this prob-

[1] Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan.
[2] Yamaha Corporation, Iwata, Shizuoka 438–0192, Japan.
[a)] daichi-k@is.naist.jp

lem, SNMF has been proposed [4], [5]. In particular, a penalized SNMF (PSNMF), which imposes a penalty term to force supervised bases and other bases to become uncorrelated with each other, achieves good performance [5]. PSNMF consists of two processes, a priori training and observed signal separation, as described below in detail.

2.1.1 Training process of supervision

In PSNMF, as the supervision training process, a priori spectral patterns (bases) should be trained in advance to achieve source 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

$$\boldsymbol{Y}_{\text{target}} \simeq \boldsymbol{F}\boldsymbol{Q}, \tag{1}$$

where $\boldsymbol{Y}_{\text{target}}(\in \mathbb{R}_{\geq 0}^{\Omega \times T_s})$ is an amplitude spectrogram of the specific signal for training, $\boldsymbol{F}(\in \mathbb{R}_{\geq 0}^{\Omega \times K})$ is a nonnegative matrix that involves bases of the target signal as column vectors, and $\boldsymbol{Q}(\in \mathbb{R}_{\geq 0}^{K \times T_s})$ is a nonnegative matrix that corresponds to the activation of each basis of $\boldsymbol{F}$. In addition, Ω is the number of frequency bins, T_s is the number of frames of the training signal, and K is the number of bases. Therefore, the basis matrix $\boldsymbol{F}$ constructed by Eq. (1) is used for the supervision of the target instrumental signal.

2.1.2 Signal separation process

The following equation represents the decomposition of PSNMF using the trained supervision matrix $\boldsymbol{F}$:

$$\boldsymbol{Y} \simeq \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}, \tag{2}$$

where $\boldsymbol{Y}(\in \mathbb{R}_{\geq 0}^{\Omega \times T})$ is an observed spectrogram, $\boldsymbol{G}(\in \mathbb{R}_{\geq 0}^{K \times T})$ is an activation matrix that corresponds to $\boldsymbol{F}$, $\boldsymbol{H}(\in \mathbb{R}_{\geq 0}^{\Omega \times L})$ represents the residual spectral patterns that cannot be expressed by $\boldsymbol{F}\boldsymbol{G}$, and $\boldsymbol{U}(\in \mathbb{R}_{\geq 0}^{L \times T})$ is an activation matrix that corresponds to $\boldsymbol{H}$. Moreover, T is the number of frames of the observed signal and L is the number of bases of $\boldsymbol{H}$. In PSNMF, the matrices $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ are optimized under the condition that $\boldsymbol{F}$ is known in advance. Hence, ideally, $\boldsymbol{F}\boldsymbol{G}$ represents the target instrumental components and $\boldsymbol{H}\boldsymbol{U}$ represents the components other than the target sounds after the decomposition.

2.1.3 Cost function

In the decomposition of PSNMF, a cost function is defined using some measures of the distance between $\boldsymbol{Y}$ and $\boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}$ as

$$\mathcal{J}_{\text{SNMF}} = \mathcal{D}(\boldsymbol{Y}|\boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}) + \mu\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|_{\text{Fr}}^2, \tag{3}$$

where $\mathcal{D}(\cdot|\cdot)$ is an arbitrary distance function, e.g., Itakura-Saito divergence (*IS-divergence*), generalized Kullback-Leibler divergence (*KL-divergence*), or the Euclidean distance (*EUC-distance*). In this study, we use EUC-distance in the cost function. In addition, μ is the weighting parameter for the penalty term and $\|\cdot\|_{\text{Fr}}$ represents the Frobenius norm. This penalty term indicates that $\boldsymbol{F}$ and $\boldsymbol{H}$ are forced to become uncorrelated with each other to avoid sharing the same basis.

2.1.4 Multiplicative update rules of PSNMF

The update rules based on EUC-distance are given by

$$g_{k,t} \leftarrow g_{k,t}\frac{\sum_\omega f_{\omega,k}y_{\omega,t}}{\sum_\omega f_{\omega,k}r_{\omega,t}}, \tag{4}$$

$$h_{\omega,l} \leftarrow h_{\omega,l}\frac{\sum_t y_{\omega,t}u_{l,t}}{\sum_t u_{l,t}r_{\omega,t} + 2\mu\sum_k f_{\omega,k}\sum_{\omega'} f_{\omega',k}h_{\omega',l}}, \tag{5}$$

$$u_{l,t} \leftarrow u_{l,t}\frac{\sum_\omega h_{\omega,l}y_{\omega,t}}{\sum_\omega h_{\omega,l}r_{\omega,t}}, \tag{6}$$

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

$$r_{\omega,t} = \sum_{k'} f_{\omega,k'}g_{k',t} + \sum_{l'} h_{\omega,l'}u_{l',t}. \tag{7}$$

2.1.5 Problem of PSNMF

PSNMF can extract 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.

2.2 Directional Clustering and Its Hybrid Method with PSNMF

Decomposition methods employing directional information for the multichannel signal have also been proposed as unsupervised separation techniques [8], [9]. These methods quantize directional information via time-frequency binary masking under the assumption that the sources are completely sparse in the time-frequency domain. Such directional clustering works well, even in an underdetermined situation where the number of sources is greater than that of inputs. 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.

To solve this problem, a hybrid method that concatenates PSNMF after the directional clustering has been proposed [10]. This hybrid method can effectively extract the target instrument because the directionally clustered signal contains only few instruments. Moreover, the residual interfering signal in the same direction can be removed by PSNMF.

3. Proposed Method

3.1 Motivation and Strategy

The conventional hybrid method has a problem that the extracted signal suffers from the generation of considerable distortion. This is due to the binary masking in directional clustering. The signal in the target direction, which is obtained by directional clustering, has many spectral chasms because the assumption of sparseness in the time-frequency domain does not always hold completely. In other words, the *resolution* of the spectrogram clustered as the target-direction component is degraded by time-frequency binary masking. Figure 1 shows an example of the spectrum of a signal separated by directional clustering. The obtained spectrum has many chasms owing to the binary masking. These spectral losses may deteriorate the performance of separation because PSNMF is forced to incorrectly fit these spectral chasms using supervised bases.

To solve this problem, in this section, we propose a new

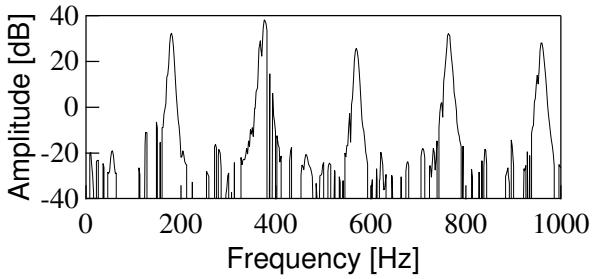

Fig. 1 Example of spectrum of signal separated by directional clustering.

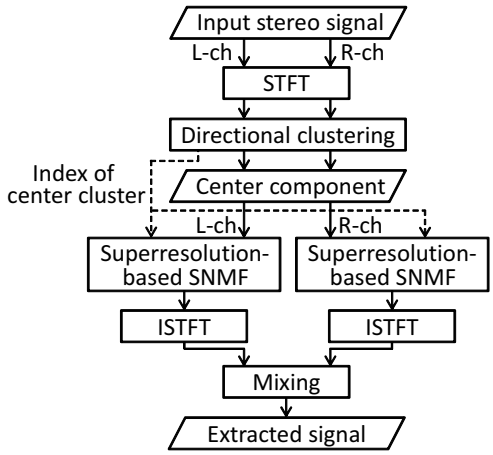

Fig. 2 Signal flow of proposed hybrid method.

superresolution-based SNMF algorithm as an alternative to the conventional PSNMF for the hybrid method (see Fig. 2). This algorithm utilizes index information determined in directional clustering. For example, if the target instrument is localized in the center cluster along with the interference, superresolution-based SNMF is only applied to the existing center components using index information. 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. In addition, the components of the target instrument lost after directional clustering can be extrapolated using the supervised bases. In other words, the resolution of the target spectrogram is recovered with the superresolution by the supervised basis extrapolation.

To illustrate the separation mechanism step by step, Fig. 3 (a) shows the configuration of source components in the stereo signal, (b) shows the separated components that are clustered around the center direction by directional clustering, and (c) shows the separated target component obtained by superresolution-based SNMF. In Fig. 3 (a), the source components are distributed in all directions with some overlapping. After directional clustering (Fig. 3 (b)), the center sources lose some of their components (i.e., the tails on both sides), and the other source components leak in the center cluster. After SNMF, the proposed algorithm restores the lost components using the supervised bases (Fig. 3 (c)).

3.2 Regularization of Basis Extrapolation

The basis extrapolation in the proposed SNMF includes an underlying problem. If the time-frequency spectra are almost unseen in the spectrogram, which means that the indexes are almost zero, a large extrapolation error may occur. Then, incorrect bases

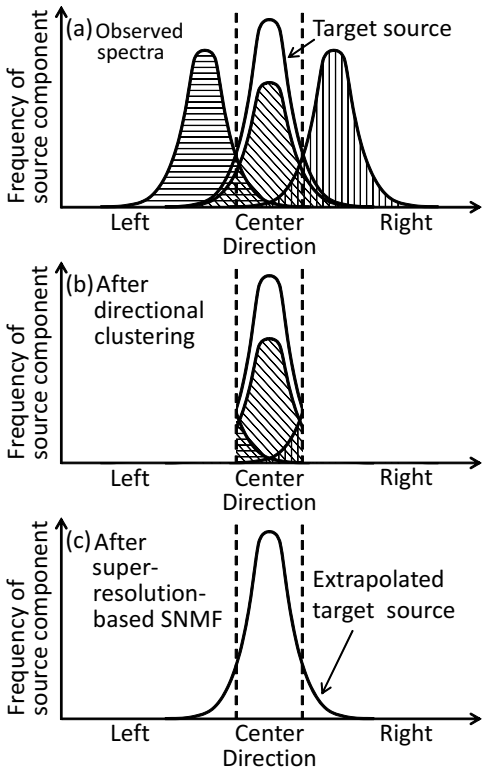

Fig. 3 Directional source distribution of (a) observed stereo signal, (b) separated components in center cluster, and (c) component separated and extrapolated by superresolution-based SNMF.

are chosen and fitted to a small number of spectral grids by incorrectly modifying the activation matrix $\boldsymbol{G}$. In the worst case, the activation matrix $\boldsymbol{G}$ contains very large values and the extracted signal is overloaded.

Figure 4 shows an example of the spectrogram that is extracted signal with basis extrapolation error at the end of signal. In this case, incorrect extrapolation occurred at the time frames that have spectral chasms as many as 99% of frequency bins. To avoid this error, we should add a new penalty term in the cost function, as described below in detail.

3.3 Cost Function of Superresolution-Based SNMF

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

$$\begin{aligned}\mathcal{J} &= \mathcal{D}(\boldsymbol{I}\circ\boldsymbol{Y}|\boldsymbol{I}\circ(\boldsymbol{F}\boldsymbol{G}+\boldsymbol{H}\boldsymbol{U})) \\ &\quad + \mu\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|^2_{\mathrm{Fr}} + \lambda\|\overline{\boldsymbol{I}}\circ(\boldsymbol{F}\boldsymbol{G})\|^2_{\mathrm{Fr}}, \end{aligned} \tag{8}$$

where μ and λ are the weighting parameters for each penalty term, $\bar{\cdot}$ represents the binary complement of each entry in the index matrix, and $\circ$ indicates the Hadamard product of matrices. Since the divergence $\mathcal{D}$ is only defined in grids whose index is one, the chasms in the spectrogram are ignored in this SNMF decomposition. In addition, the first penalty term has the same property as the conventional method Eq. (3), and the second penalty term forces the minimization of the norm $\|\boldsymbol{F}\boldsymbol{G}\|_{\mathrm{Fr}}$ in proportion

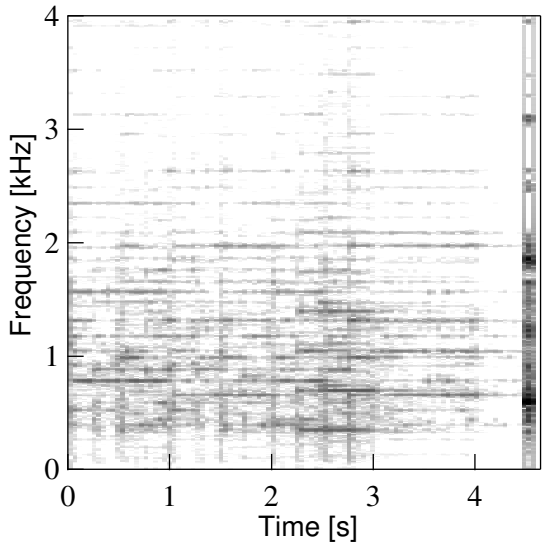

Fig. 4 Spectrogram of extracted signal with basis extrapolation error. This example incorrectly includes overloaded artifact around 4.5 sec.

to the number of zeros in the index matrix $\boldsymbol{I}$. Hence, the supervised bases are chosen so as to minimize the norm $\|\boldsymbol{FG}\|_{\mathrm{Fr}}$ to avoid the extrapolation error. In other words, this penalty term regulates the extrapolation.

3.4 Auxiliary Function for Cost Function

In this section, we derive auxiliary functions for Eq. (8) based on EUC-distance for superresolution-based SNMF. Here, we can rewrite Eq. (8) using EUC-distance as

$$\mathcal{J} = \sum_{\omega,t} i_{\omega,t}\left(y_{\omega,t}^2 + v_{\omega,t} + 2w_{\omega,t}\right) + \mu\sum_{k,l}\left(\sum_{\omega} f_{\omega,k}h_{\omega,l}\right)^2 + \lambda\sum_{\omega,t}\left(\overline{i_{\omega,t}}\sum_{k} f_{\omega,k}g_{k,t}\right)^2, \quad (9)$$

where $i_{\omega,t}$ is the entry of index matrix $\boldsymbol{I}$, which maps the values of one and zero onto the time-frequency (ω-t) region. In addition, $v_{\omega,t}$ and $w_{\omega,t}$ are given by

$$v_{\omega,t} = \left(\sum_{k} f_{\omega,k}g_{k,t}\right)^2 + \left(\sum_{l} h_{\omega,l}u_{l,t}\right)^2, \quad (10)$$

$$w_{\omega,t} = \left(\sum_{k} f_{\omega,k}g_{k,t}\right)\left(\sum_{l} h_{\omega,l}u_{l,t}\right) - y_{\omega,t}\sum_{k} f_{\omega,k}g_{k,t} - y_{\omega,t}\sum_{l} h_{\omega,l}u_{l,t}. \quad (11)$$

Since it is difficult to analytically derive the optimal $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ that minimize Eq. (9), we define an auxiliary function that represents the upper bound of $\mathcal{J}$ as described below.

First, for $v_{\omega,t}$, the upper bound function $Q^{(v_{\omega,t})}$ is defined using auxiliary variables $\alpha_{k,\omega,t} \geq 0$ and $\beta_{l,\omega,t} \geq 0$ that satisfy $\sum_k \alpha_{k,\omega,t} = 1$ and $\sum_l \beta_{l,\omega,t} = 1$. Applying Jensen's inequality to this, we have

$$v_{\omega,t} \leq \sum_{k}\frac{f_{\omega,k}^2 g_{k,t}^2}{\alpha_{k,\omega,t}} + \sum_{l}\frac{h_{\omega,l}^2 u_{l,t}^2}{\beta_{l,\omega,t}} \equiv Q^{(v_{\omega,t})}, \quad (12)$$

where the equality in (12) holds if and only if the auxiliary variables are set as

$$\alpha_{k,\omega,t} = \frac{f_{\omega,k}g_{k,t}}{\sum_{k'} f_{\omega,k'}g_{k',t}}, \quad (13)$$

$$\beta_{l,\omega,t} = \frac{h_{\omega,l}u_{l,t}}{\sum_{l'} h_{\omega,l'}u_{l',t}}. \quad (14)$$

Second, for the penalized terms (hereinafter, referred to as $\mathcal{J}^{(\mathrm{penalty})}$) in Eq. (9), the upper bound function $Q^{(\mathrm{penalty})}$ is defined using the auxiliary variables $\gamma_{\omega,k,l} \geq 0$ and $\delta_{k,\omega,t} \geq 0$ that satisfy $\sum_\omega \gamma_{\omega,k,l} = 1$, and $\sum_k \delta_{k,\omega,t} = 1$. Similarly to Eq. (12), we obtain

$$\begin{aligned}\mathcal{J}^{(\mathrm{penalty})} &= \mu\sum_{k,l}\left(\sum_{\omega} f_{\omega,k}h_{\omega,l}\right)^2 + \lambda\sum_{\omega,t}\left(\overline{i_{\omega,t}}\sum_{k} f_{\omega,k}g_{k,t}\right)^2 \\ &\leq \mu\sum_{k,l,\omega}\frac{f_{\omega,k}^2 h_{\omega,l}^2}{\gamma_{\omega,k,l}} + \lambda\sum_{\omega,t,k}\overline{i_{\omega,t}}\frac{f_{\omega,k}^2 g_{k,t}^2}{\delta_{k,\omega,t}} \\ &\equiv Q^{(\mathrm{penalty})}, \end{aligned} \quad (15)$$

where the equality in Eq. (15) holds if and only if the auxiliary variables are set as follows:

$$\gamma_{\omega,k,l} = \frac{f_{\omega,k}h_{\omega,l}}{\sum_{\omega'} f_{\omega',k}h_{\omega',l}}, \quad (16)$$

$$\delta_{k,\omega,t} = \frac{f_{\omega,k}g_{k,t}}{\sum_{k'} f_{\omega,k'}g_{k',t}}. \quad (17)$$

Finally, using Eqs. (12) and (15), we can define the upper bound function $\mathcal{J}^+$ for $\mathcal{J}$ as

$$\mathcal{J} \leq \mathcal{J}^+ = \sum_{\omega,t} i_{\omega,t}\left(y_{\omega,t}^2 + Q^{(v_{\omega,t})} + 2w_{\omega,t}\right) + Q^{(\mathrm{penalty})}. \quad (18)$$

3.5 Multiplicative Update Rules

In this section, we drive the update rules based on EUC-distance. The update rules with respect to each variable are determined by setting the gradient to zero. From $\partial\mathcal{J}^+/\partial g_{k,t} = 0$, we obtain

$$\sum_{\omega} i_{\omega,t}\left(\frac{f_{\omega,k}^2 g_{k,t}}{\alpha_{k,\omega,t}} + f_{\omega,k}\sum_{l} h_{\omega,l}u_{l,t} - y_{\omega,t}f_{\omega,k}\right) + \lambda\sum_{\omega}\overline{i_{\omega,t}}\frac{f_{\omega,k}^2 g_{k,t}}{\delta_{k,\omega,t}} = 0. \quad (19)$$

By substituting Eqs. (13), (14), (16), and (17) into Eq. (19), we can rewrite Eq. (19) as

$$\sum_{\omega} i_{\omega,t}f_{\omega,k}r_{\omega,t} + \lambda\sum_{\omega}\overline{i_{\omega,t}}f_{\omega,k}\sum_{k'} f_{\omega,k'}g_{k',t} = \sum_{\omega} i_{\omega,t}y_{\omega,t}f_{\omega,k}. \quad (20)$$

Then we can obtain the update rule by multiplying both sides of Eq. (20) by $g_{k,t}$ as follows:

$$g_{k,t} \leftarrow g_{k,t}\frac{\sum_\omega i_{\omega,t}y_{\omega,t}f_{\omega,k}}{\sum_\omega i_{\omega,t}f_{\omega,k}r_{\omega,t} + \lambda\sum_\omega \overline{i_{\omega,t}}f_{\omega,k}\sum_{k'} f_{\omega,k'}g_{k',t}}. \quad (21)$$

The update rules of the other variables based on EUC-distance are similarly obtained as follows:

$$h_{\omega,l} \leftarrow h_{\omega,l}\frac{\sum_t i_{\omega,t}y_{\omega,t}u_{l,t}}{\sum_t i_{\omega,t}u_{l,t}r_{\omega,t} + \mu\sum_k f_{\omega,k}\sum_{\omega'} f_{\omega',k}h_{\omega',l}}, \quad (22)$$

$$u_{l,t} \leftarrow u_{l,t}\frac{\sum_\omega i_{\omega,t}y_{\omega,t}h_{\omega,l}}{\sum_\omega i_{\omega,t}h_{\omega,l}r_{\omega,t}}. \quad (23)$$

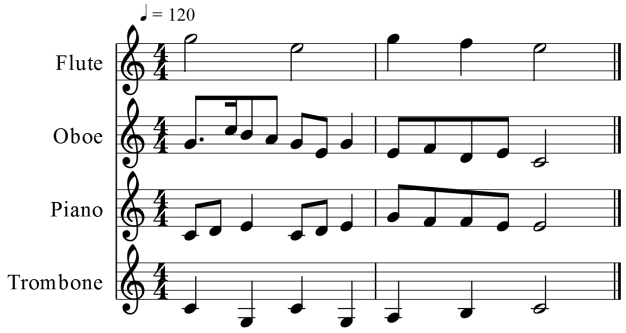

Fig. 5 Scores of each instrument.

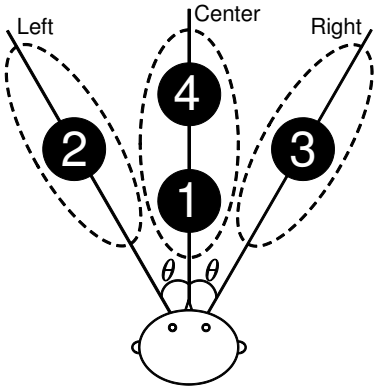

Fig. 6 Panning of four sources with sine law used in experiment. Numbered black circles represent locations of instruments in stereo format. For example, if target is Ob., No.1 is set to Ob. and Nos.2, 3, and 4 are combinations of Fl., Tb., and Pf.

4. Evaluation Experiment

4.1 Experimental Conditions

To confirm the effectiveness of the proposed algorithm, we compared five methods, namely, simple directional clustering [8], simple PSNMF [5], multichannel NMF based on IS-divergence [7], the conventional hybrid method using PSNMF after directional clustering [10], the proposed hybrid method using superresolution-based SNMF without regularization after directional clustering (hereafter, referred to as Proposed method 1), and the proposed hybrid method using superresolution-based SNMF with regularization after directional clustering (hereafter, referred to as Proposed method 2), in terms of their ability to separate simulated music signals. In this experiment, we used stereo signals containing four instruments, an oboe sound (Ob.), a flute sound (Fl.), a trombone sound (Tb.), and a piano sound (Pf.), as the signal source. The scores of each instrument are depicted in Fig. 5. These signals were artificially generated by a MIDI synthesizer, and the observed signals $\boldsymbol{Y}$ were produced by mixing four sources with an input SNR of 0 dB. The observed signal contained one source in the left and right directions and two sources in the center direction based on a sine law (see Fig. 6). The target instrument is always located in the center direction along with another interfering instrument, and the left and right sources are located at 40°. In addition, we used the same MIDI sounds of the target instruments as supervision for a priori training. The training sounds contained two octave notes that cover all notes of the target signal in the observed signal. The sampling frequency of all signals was 44.1 kHz. The spectrograms were computed using a 92-ms-long rectangular window with a 46-ms overlap shift. The number of iterations for the training was 500 and that for the separation was 400. Moreover, the number of clusters used in directional clustering was 3, the number of a priori bases was 100, and the number of bases for matrix $\boldsymbol{H}$ was 30. In this experiment, the weighting parameters μ and λ were empirically determined.

4.2 Experimental Results

We used the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and sources-to-artifacts ratio (SAR) defined in [11] as the evaluation scores. Here, the estimated signal $\hat{s}(t)$ is defined as

$$\hat{s}(t) = s_{\text{target}}(t) + s_{\text{interf}}(t) + s_{\text{artif}}(t), \tag{24}$$

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

$$\text{SDR} = 10\log_{10}\frac{\sum_t s_{\text{target}}(t)^2}{\sum_t \{e_{\text{interf}}(t) + e_{\text{artif}}(t)\}^2}, \tag{25}$$

$$\text{SIR} = 10\log_{10}\frac{\sum_t s_{\text{target}}(t)^2}{\sum_t e_{\text{interf}}(t)^2}, \tag{26}$$

$$\text{SAR} = 10\log_{10}\frac{\sum_t \left\{s_{\text{target}}(t) + e_{\text{interf}}(t)\right\}^2}{\sum_t e_{\text{artif}}(t)^2}. \tag{27}$$

SDR indicates the quality of the separated target sound, SIR indicates the degree of separation between the target and other sounds, and SAR indicates the absence of artificial distortion.

Figure 7 show the average SDR, SIR, and SAR for each method, where the four instruments are shuffled with 12 combinations. From the SDRs in Fig. 7, we can confirm that directional clustering and multichannel NMF do not have sufficient performance because they cannot discriminate the sources in the same direction. In contrast, the methods using SNMF can give better results and Proposed method 2 outperforms all other methods. Furthermore, the evaluation scores of Proposed method 1 are markedly lower than Proposed method 2. This indicates that superresolution-based SNMF in Proposed method 1 has a risk to cause the extrapolation error, whereas Proposed method 2 with regularization can mitigate such an error.

5. Conclusions

In this paper, we propose a new SNMF algorithm for the superresolution-based method to separate stereo signals and indicated an effectiveness of the regularization. The proposed algorithm utilizes index information that indicates the direction of each component in the time-frequency domain, and restores the target signal via the extrapolation of supervision bases with regularization. From the experimental results, it can be confirmed that the proposed method with regularization increases the separation performance for stereo signals compared with the conventional methods.

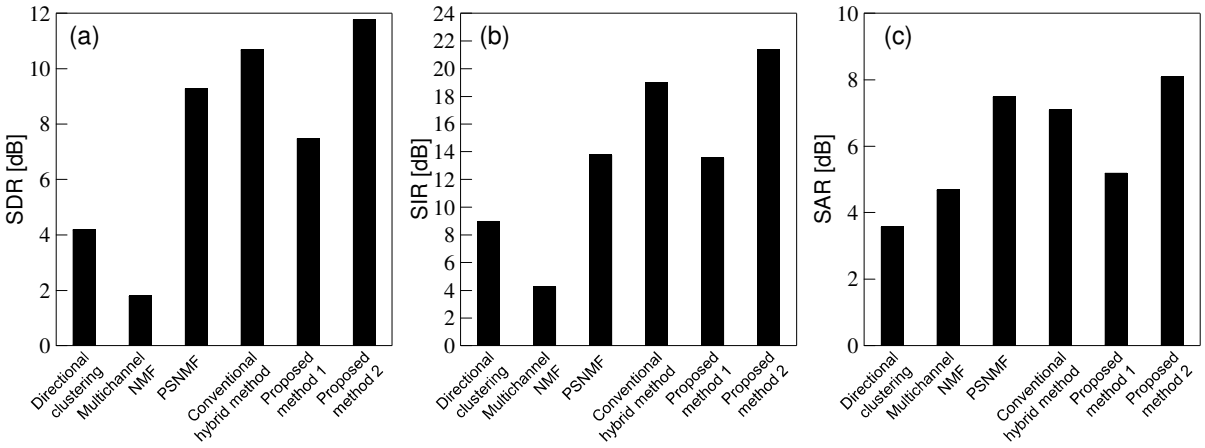

Fig. 7 Average scores when $\theta = 40°$: (a) shows SDR, (b) shows SIR, and (c) shows SAR for conventional and proposed methods.

References

[1] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," *Neural Info. Process. Syst.*, vol.13, pp.556–562, 2001.

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

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

[4] P. Smaragdis, B. Raj, M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. LVA/ICA 2010*, LNCS 6365, pp.140–148, 2010.

[5] K. Yagi, Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, "Music signal separation by orthogonality and maximum-distance constrained nonnegative matrix factorization with target signal information," *Proc. Audio Engineering Society 45th International Conference*, 2012.

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

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

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

[9] S. Araki, H. Sawada, R. Mukai, S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol.77, no.8, pp.1833–1847, 2007.

[10] Y. Iwao, H. Saruwatari, N. Kamado, K. Shikano, K. Kondo, Y. Takahashi, "Stereo music signal separation combining directional clustering and nonnegative matrix factorization," *Proc. ISSPIT*, 2012.

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