NAIST-IS-MT1251035

Master's Thesis

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

Daichi Kitamura

March 6, 2014

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Master's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
MASTER of ENGINEERING

Daichi Kitamura

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Kazushi Ikeda	(Co-supervisor)
Associate Professor Hiroshi Saruwatari	(Co-supervisor)

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

Daichi Kitamura

Abstract

In this thesis, to address a music signal separation problem, I propose a new hybrid method that concatenates directional clustering and supervised nonnegative matrix factorization (NMF) with spectrogram restoration for the purpose of the specific sound extraction from the multichannel music signal that consists of multiple instrumental sounds. Recently, a main format for obtaining musical tunes has become electronic data such as music files, which can be made available over the Internet owing to progress in information technology. Hence, users can easily obtain and edit music tunes, resulting in the active creation of new contents. According to this background, music signal separation technologies have much attention. Music signal separation is aimed to extract a specific target signal from music signals that contain multiple music instrumental sounds. Audio remixing by the users, automatic music transcription, and musical instrument education are one of the feasible music signal separation applications.

In the previous studies, music signal separation based on NMF has been a very active area of the research. Various methods using NMF have been proposed, but they remain many problems, e.g., poor convergence in update rules in NMF and lack of robustness. To solve these problems, I propose a new supervised NMF (SNMF) with spectrogram restoration and its hybrid method that concatenates the proposed SNMF after directional clustering. Via extrapolation of supervised spectral bases, this SNMF with spectrogram restoration attempts both target

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1251035, March 6, 2014.

signal separation and reconstruction of the lost target components, which are generated by preceding binary masking performed in directional clustering.

Next, I provide a theoretical analysis of basis extrapolation ability and reveal the mechanism of marked shift of optimal divergence in SNMF with spectrogram restoration and trade-off between separation and extrapolation abilities. Evaluation experiment of the separation using artificial and real-recorded music signals show the effectiveness of the proposed hybrid method.

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

Keywords:

Music signal separation, Directional clustering, Spectrogram restoration, Nonnegative matrix factorization, Supervised method.

スペクトログラム修復機能付き非負値行列因子分解と方位クラスタリングを組み合わせた音楽信号分離*

北村 大地

内容梗概

本論文では，複数の音源が多重に混合されたマルチチャネル音楽信号から頑健に目的の信号成分を分離する手法の確立を目指し，方位クラスタリングとスペクトログラム修復機能付き教師あり非負値行列因子分解 (NMF) を組み合わせた新しい音楽信号分離手法を提案する．近年，音楽メディアは電子ファイルとして供給されインターネットを通じて配信される機会が増加している．そのため，ユーザが既存の音楽メディアを自由に編集する等の能動的な創作活動が盛んになっている．このような背景から，音楽信号を対象とした信号分離手法が広く注目を集めており世界中で盛んに研究されている．この音楽信号分離技術は，複数の楽器が多重に混合された音楽信号の中から特定の楽器音を分離・抽出することを目的としており，オーディオリミックス，自動採譜，楽器演奏法のための音楽教育といった様々な応用先が考えられる．

これまでの研究において，NMF を用いた音楽信号分離技術が非常に高い注目を集めており，様々な改良手法が提案されている．しかしながら，いずれの手法においても多くの問題があり，頑健かつ高精度に目的音を分離する手法は未だ提案されていないのが現状である．そこで本研究では，新たにスペクトログラム修復機能付き教師あり NMF (SNMF) を提案し，方位クラスタリングと組み合わせることによって頑健かつ高精度に目的音を分離するマルチチャネル信号分離手法を提案する．新たに提案するスペクトログラム修復機能付き SNMF は，信号の分離と同時に，前段処理の方位クラスタリングによって失われた目的音成分の復元を教師スペクトル基底の外挿によって実現する．すなわち，前段処理によって傷

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文, NAIST-IS-MT1251035, 2014 年 3 月 6 日.

ついた混合信号のスペクトログラムを，後段の SNMF の教師スペクトル基底の外挿によって復元することができる．

次に，教師スペクトル基底の外挿能力について，信号の生成モデルに基づく理論解析を行い，基底外挿における最適なダイバージェンス規範を示し，スペクトログラム修復機能付き SNMF 特有の最適ダイバージェンスシフトのメカニズムを明らかにする．加えて，ダイバージェンス規範に対する分離能力と外挿能力のトレードオフを理論的に示し，人工及び実録音音楽信号を用いた音楽信号分離実験において解析結果と同様の現象が現れることを確認する．

さらに，明らかにされた外挿能力理論に基づいて異なるダイバージェンス規範をフレームごとに切り替える多重ダイバージェンス型ハイブリッド手法を提案する．このダイバージェンスのダイバーシチを実装した提案ハイブリッド手法はいかなる空間的配置の入力信号に対しても常に最高の分離性能を達成し，提案手法の頑健性をさらに向上させることができる．本手法の有効性は，音楽信号分離実験によって確認される．

キーワード

音楽信号分離，非負値行列因子分解，方位クラスタリング，スペクトログラム修復，教師あり手法.

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

In recent years, the main format for obtaining musical tunes has become electronic data such as music files, which can be made available over the Internet owing to progress in information technology. Hence, users can easily obtain and edit music tunes, resulting in the active creation of new contents. These consumers' activities have rapidly increased in the past few years with the expansion of social networking services and video-sharing websites. However, it is still difficult to edit a specific instrumental signal in general music tunes containing many instruments because almost all the commercially available music data are mixed down, and consumers cannot obtain each solo-played instrumental signal in advance. Such audio editing of each sound source will enable us to engage in new activities based on the appreciation of musical tunes and can be applied to many valuable techniques including audio remixing by users [1, 2], musical instrument education, 3D audio reproduction [3], and automatic music transcription [4, 5] (see Fig. 1). With this background, music signal separation technologies have attracted considerable interest and been intensively studied [6, 7, 8, 9, 10] in recent years. However, it remains difficult to freely extract a specific music signal, particularly in the case of instruments that belong to the same family.

Signal separation can be classified into *overdetermined* and *underdetermined* problems. In the former situation, the number of channels is greater than the number of sound sources, and many techniques have been studied and proposed for overdetermined signal separation [11, 12, 13]. However, such separation techniques cannot be applied to the above-mentioned music signal separation problem because almost all musical tunes are provided in a stereo format and the number of sources is greater than two. Therefore, techniques for underdetermined separation are required and should be used to achieve music signal separation.

1.2 Prior works

As a means of addressing underdetermined signal separation, in recent years, nonnegative matrix factorization (NMF) [14], which is a type of sparse representation

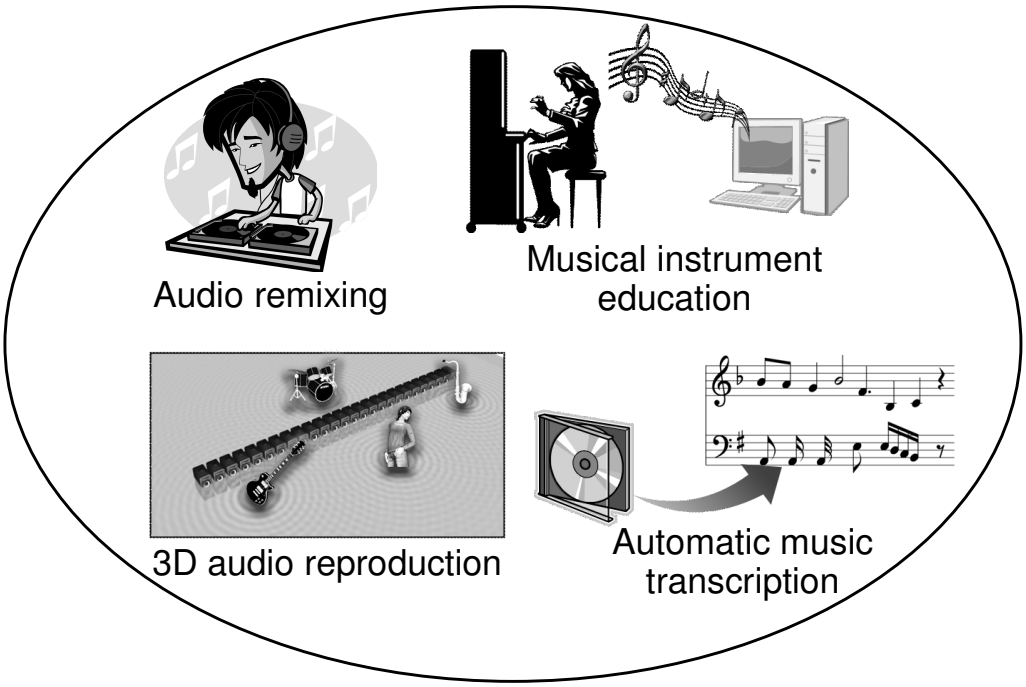

Figure 1. Applications of music signal separation.

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 [15, 16, 17, 18, 19]. 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) [20] and its improved method, penalized SNMF (PSNMF) [21, 23, 22], have been proposed. These methods include a priori training, which requires some sound samples of a target instrument, and separate the target signal using supervised bases. PSNMF 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, such as more realistic musical tunes, the source 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 [24, 25]. This method is a natural extension of NMF for a stereo or multichannel signal and is a unified method that addresses the spatial and spectral separation problems simultaneously. However, such unsupervised separation is a difficult problem, even if the signal has multichannel components, because the decomposition is underspecified. Hence, these algorithms involve strong dependence on initial values and lack robustness. For multichannel signal separation, directional clustering has also been proposed as an unsupervised method [26, 27, 28]. 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 only the directional information. To cope with this problem, a hybrid method for multichannel signal separation, which concatenates PSNMF after directional clustering, has been proposed [29]. 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, which mean spectral holes in the spectrogram. This results in the cascaded SNMF being forced to incorrectly mimic such artificial spectral chasms.

In summary, no effective technique has yet been proposed for separating the target source from a multichannel signal with high accuracy and satisfactory robustness. Therefore, attempts should be made to develop an effective algorithm for underdetermined signal separation. Such a robust signal separation method for multichannel signals will be applicable to not only music signals but also speech signals recorded by a microphone array to enhance the speech and suppress interfering noise.

1.3 Scope of thesis

To achieve high-quality music signal separation with robustness, in this thesis, I propose a new hybrid method that concatenates a new SNMF algorithm and an unsupervised multichannel signal separation method. In addition, I also provide a mathematical analysis for optimizing the proposed hybrid method. Figure 2

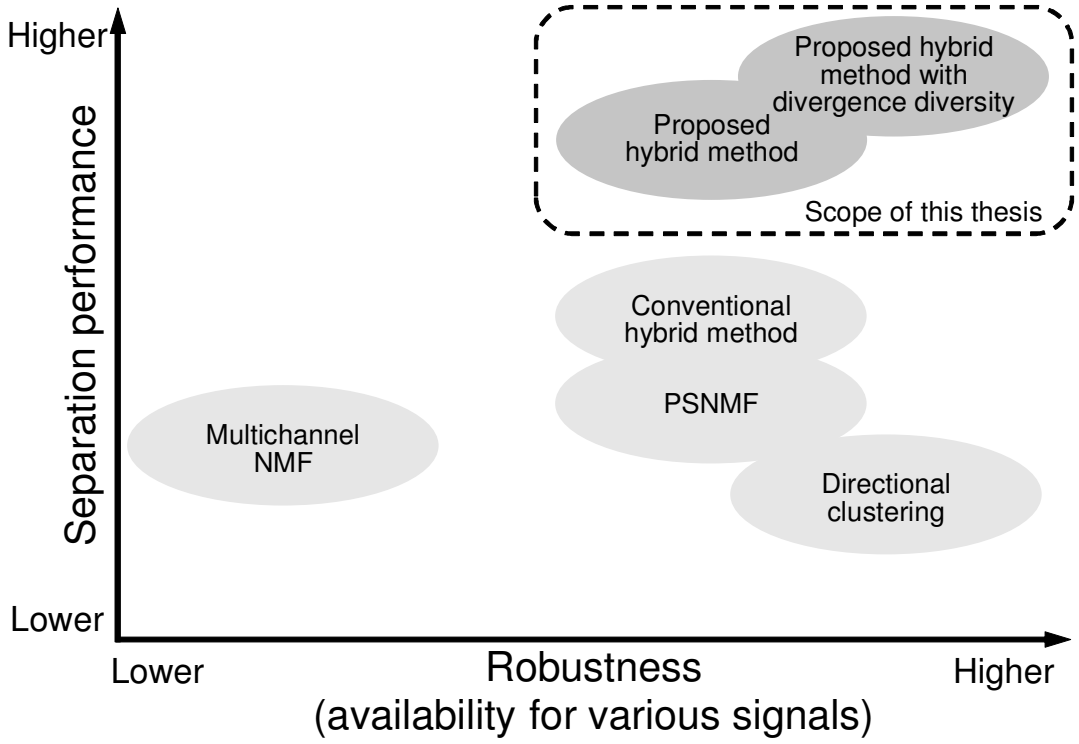

Figure 2. Relation between conventional methods and proposed hybrid methods.

depicts the relation between the conventional methods and the proposed hybrid methods.

The hybrid method divides the stereo music signal separation problem into two stages, namely, *spatial separation* and *spectral separation*. The spatial separation utilizes binary masking, which is performed by the directional clustering technique, in the time-frequency domain. A clustering approach using spatial information is a common strategy used for multichannel signal separation because it works well even in underdetermined situations. Then, in the spectral separation stage, a new SNMF algorithm is applied to separate the signals in the same direction. In addition, this SNMF algorithm improves the sound quality of the target signal, which is deteriorated by the preceding binary masking performed in directional clustering. Therefore, the proposed hybrid method is a divide-and-conquer method that utilizes suitable decompositions in each separation problem

and achieves robust multichannel signal separation with less sensitivity to the initial values.

In the conventional hybrid method [29], the spectral chasms generated by binary masking degrade the sound quality of the separated target signal because PSNMF is concatenated directly after directional clustering. To solve this problem, I propose a new SNMF with spectrogram restoration. By utilizing index information generated from binary masking, the proposed SNMF regards the spectral chasms as *unseen* observations, and finally reconstructs the target signal components via spectrum extrapolation using supervised bases. In other words, this SNMF can be categorized as a *inpainting* because the deteriorated spectrogram resulting from the preceding binary masking can be recovered. Note that an SNMF-based extrapolation technique for acoustic signals has been proposed as a means of expanding the acoustic signal bandwidth [30]. However, this method cannot be applied to signal separation.

The proposed SNMF with spectrogram restoration attempts both signal *separation* and basis *extrapolation* using the supervised bases. In previous studies, the analysis of the optimal divergence criterion in SNMF has only been discussed for signal separation [21, 22, 31], and the issue of the optimal divergence criterion in SNMF for basis extrapolation has not been addressed. Therefore, in this thesis, I analyze the ability of basis extrapolation for each divergence criterion in SNMF.

1.4 Outline of thesis

The thesis is organized as follows. First, I describe related works on single-channel and multichannel signal separation methods in Sect. 2. In this section, an overview of NMF is also given. In Sect. 3, I propose a new SNMF with spectrogram restoration and derive its update rules for optimization. Also, the relation between the extrapolation ability and the divergence criterion in SNMF is clarified by theoretical analysis based on a signal generation model to find the optimal criterion for SNMF with spectrogram restoration. In addition, the efficacy of the proposed hybrid method with the proposed SNMF is confirmed experimentally for musical signal separation. On the basis of the above-mentioned findings, in Sect. 4, I propose a new method for switching the divergence criterion in SNMF with spectrogram restoration to adapt to various types of input signals

and to separate the target signal robustly. The robustness of the proposed method is confirmed by experimental evaluations. Finally, I summarize the contributions of this thesis and provide suggestions for future work in Sect. 5.

2. Conventional Signal Separation Methods

2.1 Introduction

In this section, I describe conventional music signal separation methods and their problems. In recent years, many types of signal separation methods have been proposed and studied. These methods are roughly classified into single-channel and multichannel signal separation algorithms. The former method attempts the underdetermined separation using some constraints derived from the property of the target signal. The latter method uses a spatial cue, which is obtained as difference between channels, as a directional information and separates the target signals. Then, in this section, I review commonly used signal separation methods, PSNMF, Directional clustering, and Multichannel NMF.

First, I outline conventional single-channel signal separation methods in Sect. 2.2. Next, I give a brief review of multichannel signal separation methods and its problems in Sect. 2.3. Finally, Sect. 2.4 concludes this section.

2.2 Conventional single-channel signal separation methods

2.2.1 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 (or a power) spectrogram for applying NMF to the acoustic signal; $\boldsymbol{V}(\in \mathbb{R}_{\geq 0}^{M\times D})$ is often called the *basis matrix*, which includes bases (frequently-appearing spectral patterns in $\boldsymbol{X}$) as column vectors; and $\boldsymbol{W}(\in \mathbb{R}_{\geq 0}^{D\times N})$ is often called the *activation matrix*, which involves activation information of each basis of $\boldsymbol{V}$. In addition, M and N are the numbers of rows and columns of $\boldsymbol{X}$, and D is the number of bases of $\boldsymbol{V}$. Figure 3 depicts the decomposition model of NMF, where the number of bases D equals two. 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

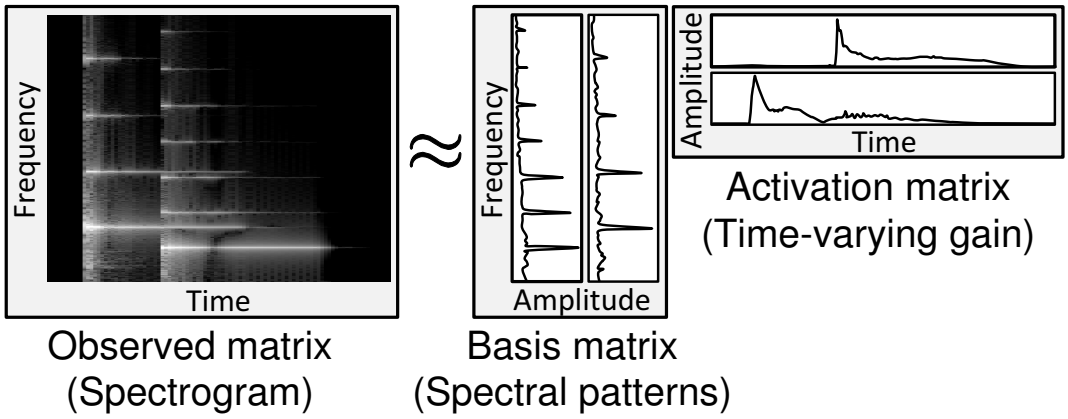

Figure 3. Decomposition model of simple NMF.

optimize the variables $\boldsymbol{V}$ and $\boldsymbol{W}$ using an arbitrary divergence between $\boldsymbol{X}$ and $\boldsymbol{VW}$. The following equation represents the cost function of NMF:

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

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, I use the following generalized divergence called β-divergence [32] in the cost function:

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

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) as shown in Figs. 6–8.

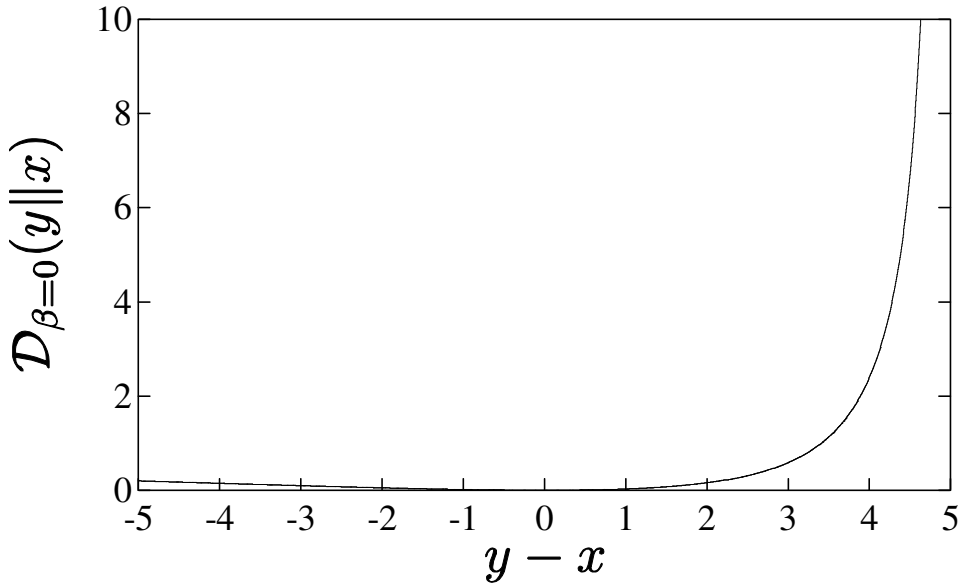

Figure 4. Variation of β-divergence function when $\beta = 0$.

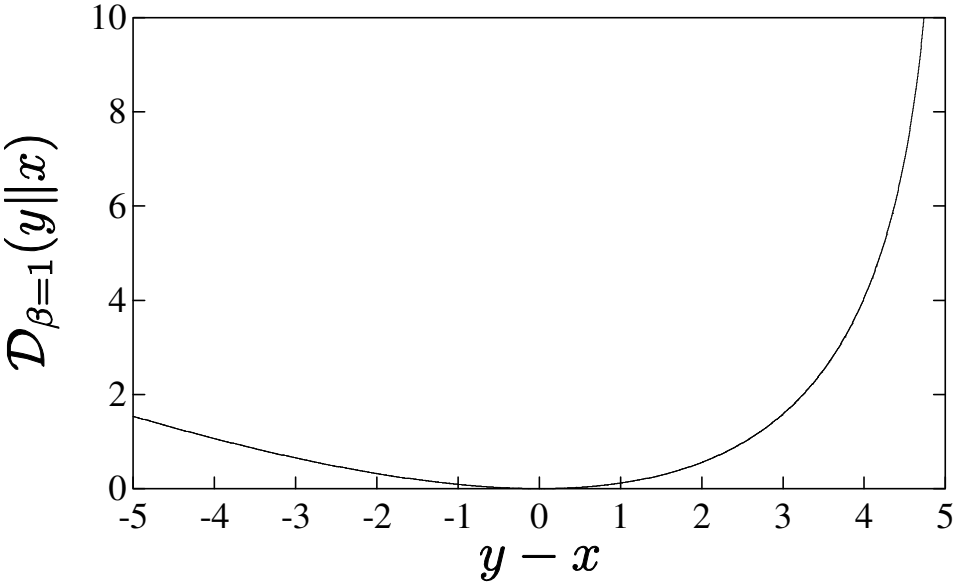

Figure 5. Variation of β-divergence function when $\beta = 1$.

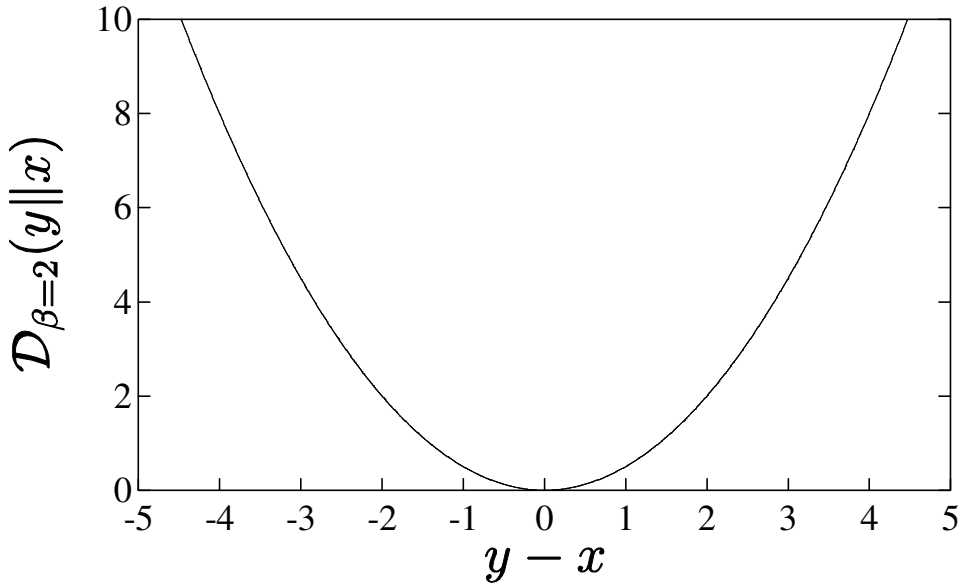

Figure 6. Variation of β-divergence function when $\beta = 2$.

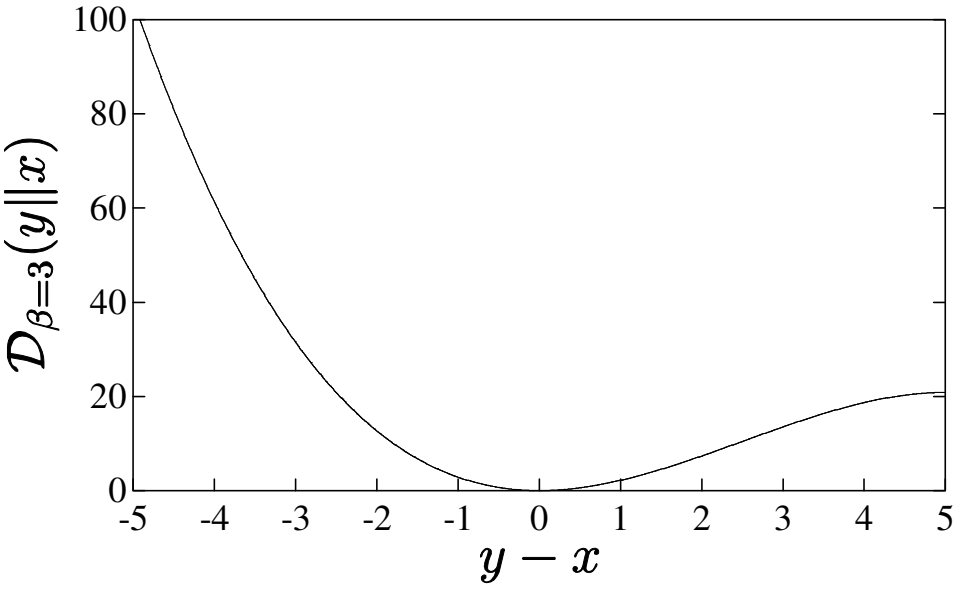

Figure 7. Variation of β-divergence function when $\beta = 3$.

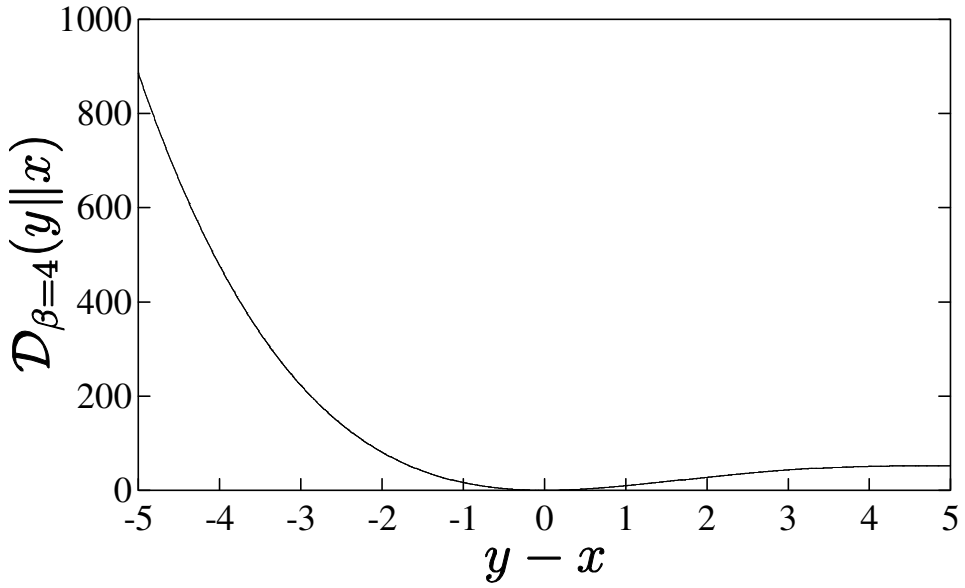

Figure 8. Variation of β-divergence function when $\beta = 4$.

The multiplicative update rules for $\boldsymbol{V}$ and $\boldsymbol{W}$ that minimize the cost function based on β-divergence are given by [33]

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

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

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

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

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

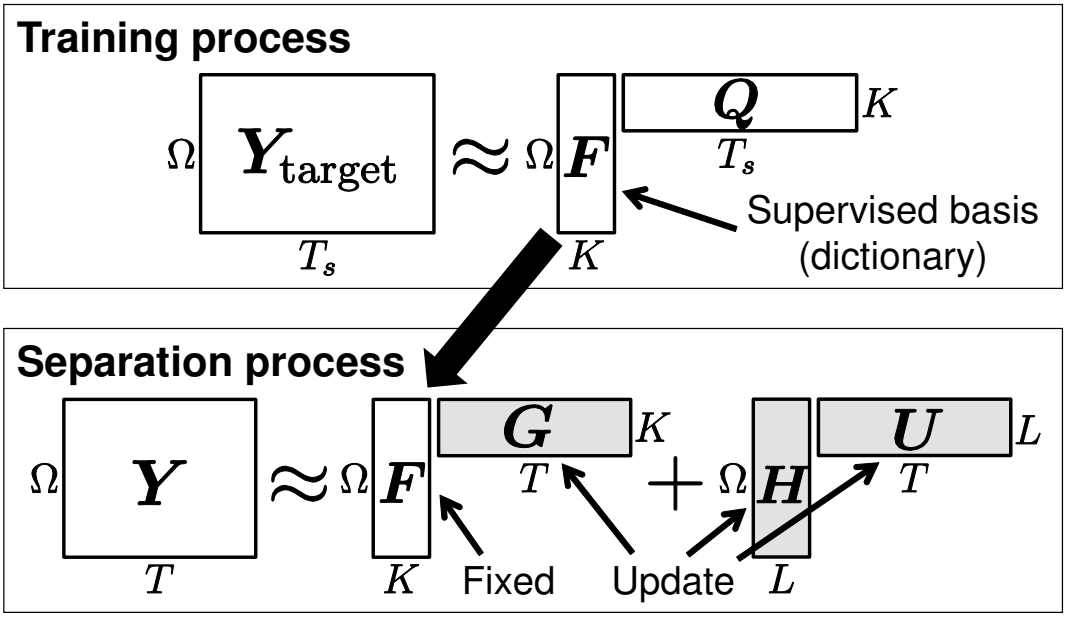

Figure 9. Decomposition model in each process of SNMF.

2.2.2 SNMF and PSNMF

The signal separation using NMF is achieved by extracting only the target spectral bases. However, such unsupervised approaches have difficultly in clustering the decomposed spectral patterns into a specific target instruments. Furthermore, each basis may be forced to include a multi-instrumental spectral pattern. To solve this problem, SNMF [20] and its improved method, PSNMF, have been proposed [21, 22]. These supervised scheme consists of two processes, namely, a priori training and observed signal separation as shown in Fig. 9.

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

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

where $\boldsymbol{Y}_{\text{target}} (\in \mathbb{R}_{\geq 0}^{\Omega \times T_s})$ is an amplitude spectrogram of the specific instrumental signal for training, $\boldsymbol{F} (\in \mathbb{R}_{\geq 0}^{\Omega \times K})$ is a nonnegative matrix that involves bases of the

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

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

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

where $\boldsymbol{Y}(\in \mathbb{R}_{\geq 0}^{\Omega \times T})$ is an observed spectrogram, $\boldsymbol{G}(\in \mathbb{R}_{\geq 0}^{K \times T})$ is an activation matrix that corresponds to $\boldsymbol{F}$, $\boldsymbol{H}(\in \mathbb{R}_{\geq 0}^{\Omega \times L})$ is the residual spectral patterns that cannot be expressed by $\boldsymbol{F}\boldsymbol{G}$, 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 SNMF, the matrices $\boldsymbol{G}$, $\boldsymbol{H}$, and $\boldsymbol{U}$ are optimized under the condition that $\boldsymbol{F}$ is known in advance. Hence, ideally, $\boldsymbol{F}\boldsymbol{G}$ represents the target instrumental components, and $\boldsymbol{H}\boldsymbol{U}$ represents other different components from the target sounds after the decomposition. The cost function for (8) is defined as

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

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

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

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

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

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

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

However, this SNMF incurs a risk of degrading the separation performance owing to the simultaneous generation of similar spectral patterns in the supervised basis matrix $\boldsymbol{F}$ and other basis matrix $\boldsymbol{H}$ (referred to as basis sharing problem). This is because the cost function in SNMF is defined as the divergence between the observed and reconstructed matrices, and unique decomposition is not guaranteed. To solve this problem, PSNMF has been proposed [21, 22]. PSNMF employs a penalty term in the cost function to force the other bases to become as different as possible from the supervised bases.

The cost function of PSNMF with orthogonality penalty is defined as follows:

$$\mathcal{J}_{\mathrm{PSNMF1}} = \mathcal{D}_{\beta_{\mathrm{NMF}}}(\boldsymbol{Y} \| \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}) + \mu_1 \| \boldsymbol{F}^{\mathrm{T}}\boldsymbol{H} \|_{\mathrm{Fr}}^2, \tag{14}$$

where the conditions $\sum_\omega f_{\omega,k} = 1$ and $\sum_\omega h_{\omega,l} = 1$ are applied, μ_1 is a weighting parameter for the penalty term, and $\| \cdot \|_{\mathrm{Fr}}$ indicates the Frobenius norm. The minimization of the second term in (14) corresponds to the maximization of orthogonality between $\boldsymbol{F}$ and $\boldsymbol{H}$. The update rule for $\boldsymbol{H}$, which minimizes the cost function (14), is given by

$$h_{\omega,l} \leftarrow h_{\omega,l} \left(\frac{\sum_t y_{\omega,t} u_{l,t} z_{\omega,t}^{\beta_{\mathrm{NMF}}-2}}{\sum_t u_{l,t} z_{\omega,t}^{\beta_{\mathrm{NMF}}-1} + 2\mu_1 \sum_k f_{\omega,k} \sum_{\omega'} f_{\omega',k} h_{\omega',l}} \right)^{\varphi(\beta_{\mathrm{NMF}})}. \tag{15}$$

The update rules for $\boldsymbol{G}$ and $\boldsymbol{U}$ are the same as (11) and (12).

Also, maximum-divergence penalty, which maximizes all divergence combinations between the supervised bases in $\boldsymbol{F}$ and the other bases in $\boldsymbol{H}$, has been proposed as another means of preventing the basis sharing problem. The cost function of PSNMF with maximum-divergence penalty is defined as follows:

$$\mathcal{J}_{\mathrm{PSNMF2}} = \mathcal{D}_{\beta_{\mathrm{NMF}}}(\boldsymbol{Y} \| \boldsymbol{F}\boldsymbol{G} + \boldsymbol{H}\boldsymbol{U}) + \mu_2 \exp\left(-\frac{1}{\lambda_{\mathrm{m}}} \sum_{k,l,\omega} \mathcal{D}_{\beta_{\mathrm{m}}} \left(f_{\omega,k} \| h_{\omega,l} \right) \right), \tag{16}$$

where μ_2 and λ_{m} are the weighting and sensitivity parameters, respectively. Here, exponentiation is applied to make the penalty term nonnegative.

The update rule for $\boldsymbol{H}$, which minimizes the cost function (16), is given by

$$h_{\omega,l} \leftarrow h_{\omega,l} \left(\frac{\lambda_{\mathrm{m}} \sum_t y_{\omega,t} u_{l,t} z_{\omega,t}^{\beta_{\mathrm{NMF}}-2} + \mu_2 h_{\omega,l}^{\beta_{\mathrm{m}}-1} C_{\beta_{\mathrm{m}}}}{\lambda_{\mathrm{m}} \sum_t u_{l,t} z_{\omega,t}^{\beta_{\mathrm{NMF}}-1} + \mu_2 h_{\omega,l}^{\beta_{\mathrm{m}}-2} C_{\beta_{\mathrm{m}}} \sum_k f_{\omega,k}} \right)^{\varphi(\beta_{\mathrm{NMF}})}, \tag{17}$$

where

$$C_{\beta_{\mathrm{m}}} = \exp\left(-\frac{1}{\lambda_{\mathrm{m}}}\sum_{k,l,\omega}\mathcal{D}_{\beta_{\mathrm{m}}}\left(f_{\omega,k}\|h_{\omega,l}\right)\right). \tag{18}$$

The update rules for $\boldsymbol{G}$ and $\boldsymbol{U}$ are the same as (11) and (12).

By imposing these penalty terms on the cost function, we can prevent the basis sharing problem and separate the target signal with high accuracy. The separation performance of PSNMF with orthogonality penalty and PSNMF with maximum-divergence penalty are almost the same [22]. In this thesis, hereafter, I use the orthogonality penalty and its update rules.

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.

2.3 Conventional multichannel signal separation methods

2.3.1 Directional clustering

Decomposition methods employing directional information for the multichannel signal have also been proposed as unsupervised separation techniques [26, 27, 28]. These methods quantize directional information via time-frequency binary masking under the assumption that the sources are completely sparse (double disjoint) in the time-frequency domain. Figure 10 shows the separation algorithm of directional clustering. The target source in the center direction is separated by hard clustering method, which corresponds to the binary masking 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.

2.3.2 Hybrid method of directional clustering and PSNMF

To separate the sources in the same direction, a hybrid method that concatenates PSNMF after directional clustering has been proposed [29]. Figure 11 indicates the signal flow of the hybrid method. This hybrid method can effectively extract the target instrument because the directionally clustered signal contains only few instruments. Moreover, the residual interfering signal in the same direction can be removed by PSNMF.

However, this 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 12 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.

2.3.3 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 [24, 25]. These algorithms employ Hermitian positive definite matrix that models the spatial property of each NMF basis and each frequency bin. Therefore, multichannel NMF utilizes a frequency-wise transfer function between signal source and microphone as a cue for basis clustering. However, such unsupervised separation is a difficult problem, even if the signal has multichannel components, because the decomposition is underspecified. Hence, these algorithms involve strong dependence on initial values and lack robustness.

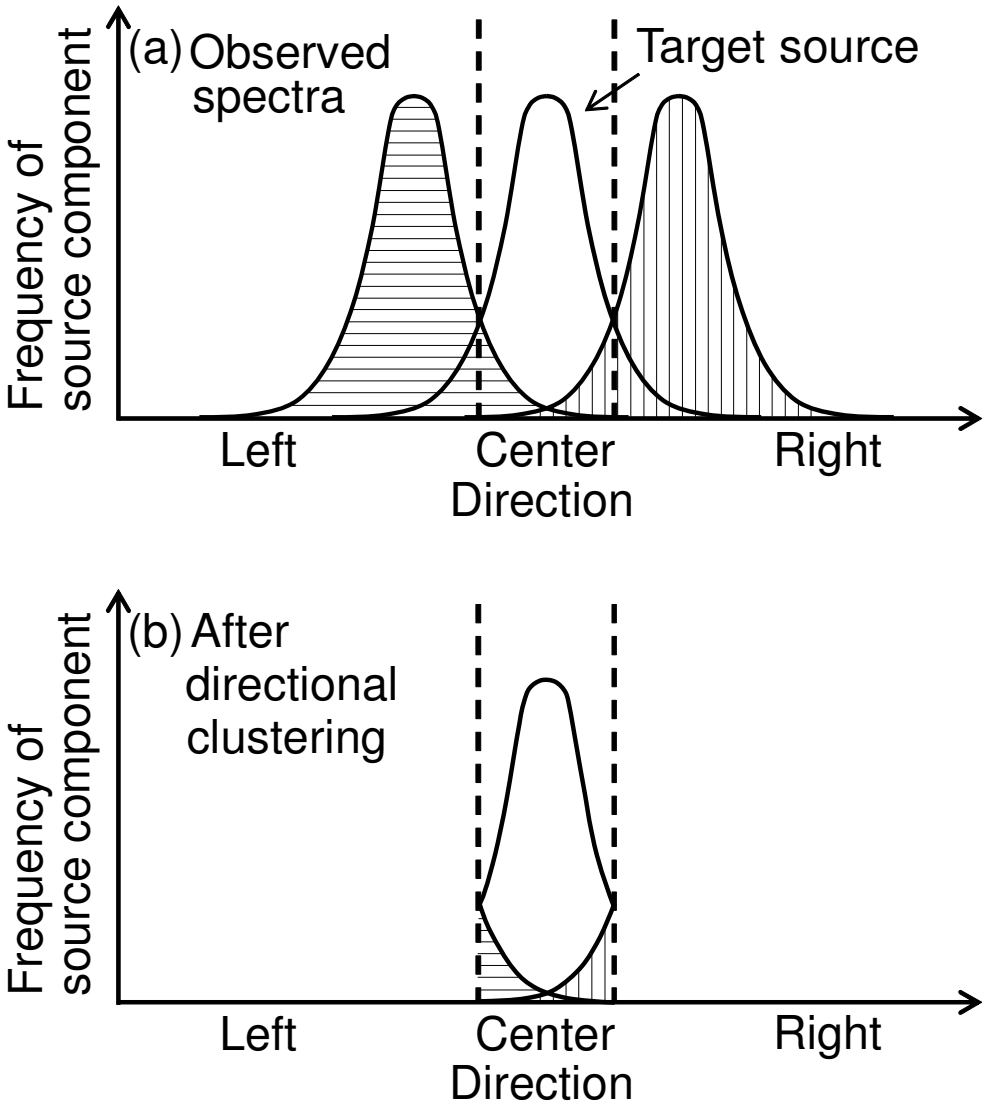

Figure 10. Directional source distribution of (a) observed stereo signal, (b) separated target components in the center cluster.

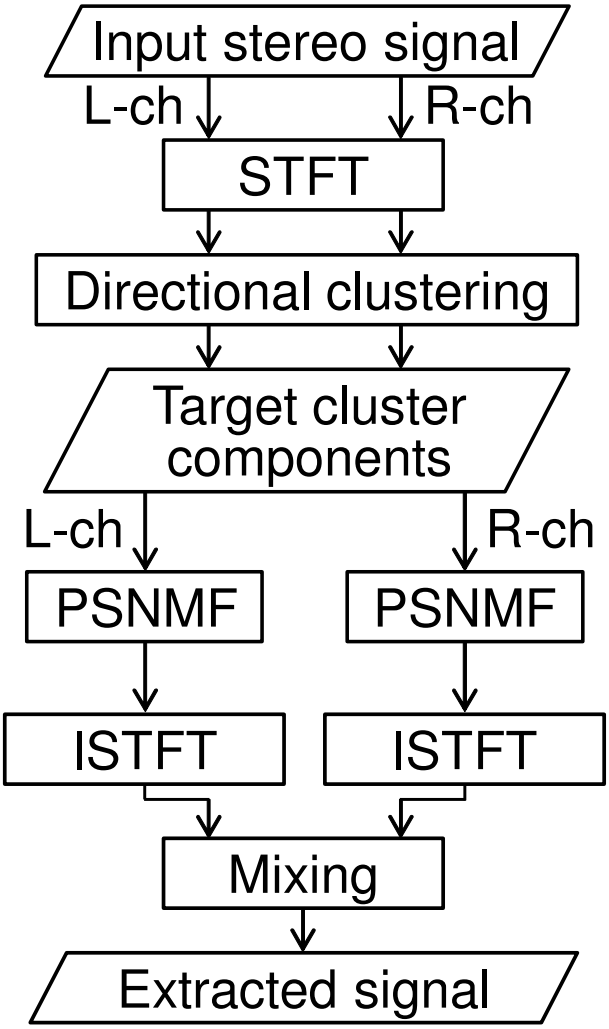

Figure 11. Signal flow of conventional hybrid method; PSNMFs are cascaded after stereo output of directional clustering.

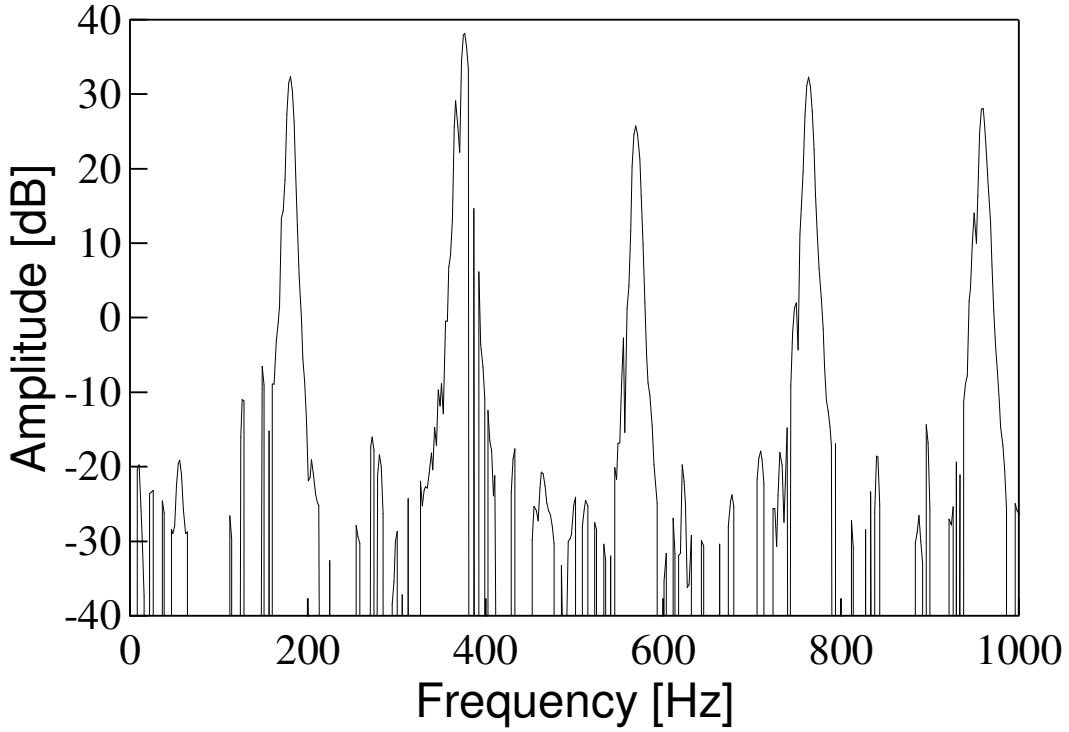

Figure 12. Example of spectrum of signal separated by directional clustering.

2.4 Conclusion

In this section, conventional single-channel signal separation methods were denoted. Next, conventional multichannel signal separation methods were reviewed. Since each method has its own drawback, I propose a new hybrid method of directional clustering and a new SNMF with spectrogram restoration in the next section to cope with the problems effectively.

3. SNMF with Spectrogram Restoration and Its Hybrid Method

3.1 Introduction

In the previous section, I described two types of conventional signal separation methods and the problems of each method. To solve these problems, in this section, I propose a new algorithm of SNMF with spectrogram restoration and its hybrid method as a multichannel signal separation method.

First, I describe a strategy and a derivation of update rules for SNMF with spectrogram restoration in Sect. 3.2. Second, theoretical analysis of basis extrapolation ability based on generation model is shown in Sect. 3.3. Third, I compare separation performance of the proposed method and the other conventional methods via some experiments in Sect. 3.4 for a validation of the proposed method. Finally, Sect. 3.5 concludes this section.

3.2 SNMF with spectrogram restoration

3.2.1 Motivation and strategy

The separated signal by the conventional hybrid method described in Sect. 2.3.2 suffers from the generation of considerable distortion owing to the binary masking in directional clustering. To solve this problem, in this section, I propose a new SNMF with spectrogram restoration as an alternative to the conventional PSNMF for the hybrid method.

Figure 13 shows a signal flow of the proposed hybrid method that includes SNMF with spectrogram restoration. The algorithm of SNMF with spectrogram restoration utilizes index information determined in directional clustering. For example, if the target instrument is localized in the center cluster along with the interference, SNMF is only applied to the existing center components using index information (active binary mask). Therefore, the spectrogram of the target instrument is reconstructed using more matched bases because spectral chasms are treated as *unseen*, and these chasms have no impact on the cost function in SNMF with spectrogram restoration. In addition, the components of the target

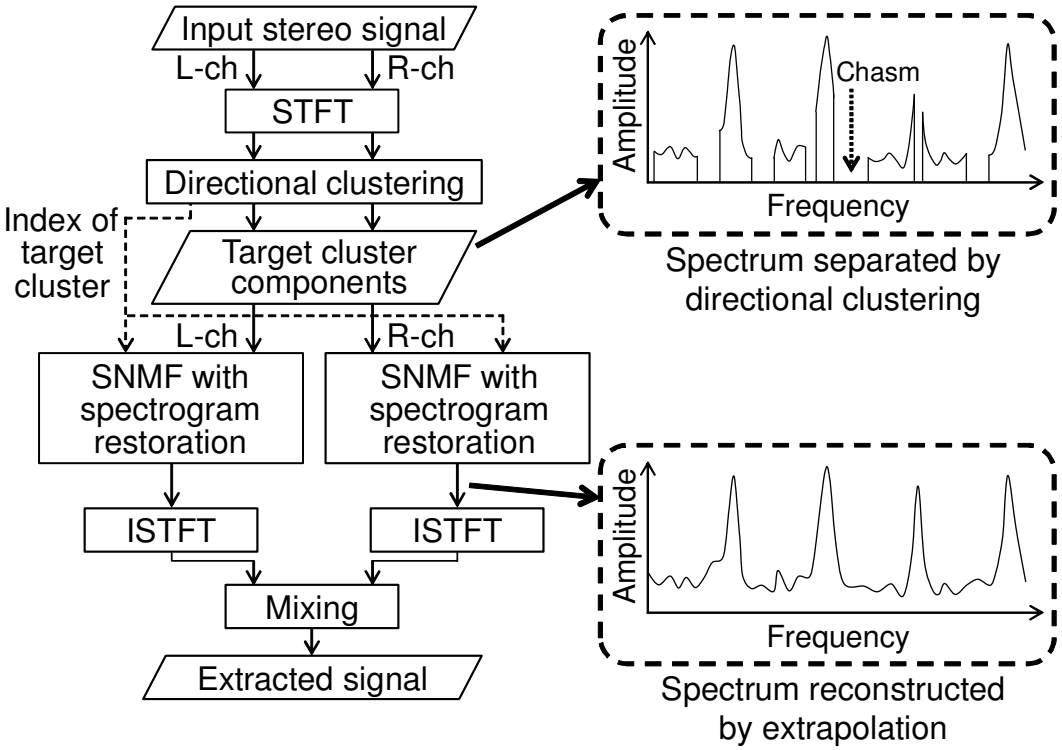

Figure 13. Signal flow of proposed hybrid method; SNMF with spectrogram restoration concatenates after directional clustering.

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

To illustrate the separation mechanism step by step, Fig. 14 (a) shows the configuration of source components in the stereo signal, (b) shows the separated components that are clustered around the center direction by directional clustering, and (c) shows the separated target component obtained by SNMF with spectrogram restoration. In Fig. 14 (a), the source components are distributed in all directions with some overlapping. After directional clustering (Fig. 14 (b)), the center sources lose some of their components (i.e., the tails on both sides), and the other source components leak in the center cluster. After SNMF with spectrogram restoration, the proposed algorithm restores the lost components using the supervised bases (Fig. 14 (c)).

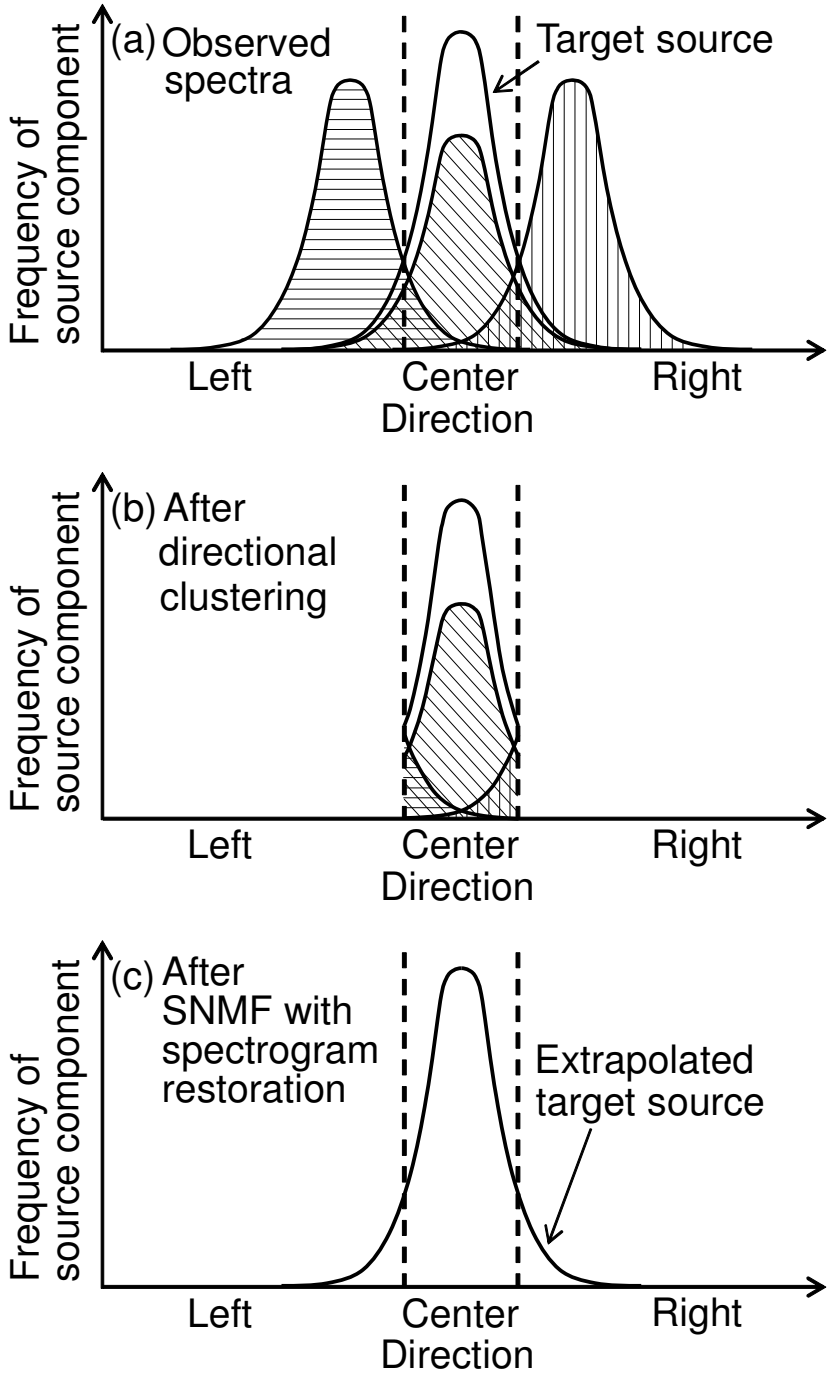

Figure 14. Directional source distribution of (a) observed stereo signal, (b) separated components in center cluster, and (c) component separated and extrapolated by spectrogram restoration.

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

3.2.2 Cost function and update rules

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

$$\begin{aligned}\mathcal{J}(\Theta) = &\sum_{\omega,t} i_{\omega,t}\mathcal{D}_{\beta_{\mathrm{NMF}}}(y_{\omega,t}\|\textstyle\sum_k f_{\omega,k}g_{k,t} + \sum_l h_{\omega,l}u_{l,t}) \\ &+ \lambda\sum_{\omega,t}\overline{i_{\omega,t}}\mathcal{D}_{\beta_{\mathrm{reg}}}(0\|\textstyle\sum_k f_{\omega,k}g_{k,t}) + \mu\|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{H}\|^2_{\mathrm{F}},\end{aligned} \tag{19}$$

where $\Theta = \{\boldsymbol{G}, \boldsymbol{H}, \boldsymbol{U}\}$ is the set of objective variables, $i_{\omega,t}$ is a entry of the index matrix $\boldsymbol{I}$, λ and μ are the weighting parameters for each term, and $\bar{\cdot}$ 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 $\mathcal{D}_{\beta_{\mathrm{NMF}}}(\cdot\|\cdot)$ is only defined in grids whose index is one, the chasms in the spectrogram are ignored in this SNMF decomposition. The second term forces the minimization of the value of $\sum_k f_{\omega,k}g_{k,t}$. 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. In other words, this penalty term regulates the extrapolation. In addition, the third penalty term has the same property as that in the cost function of conventional PSNMF (14).

The update rules based on (19) are obtained by the auxiliary function approach, similarly to [33]. Here, we can rewrite the cost function (19) using β-

divergence as

$$\mathcal{J}(\Theta) = \mathcal{J}_1 + \lambda\mathcal{J}_2 + \mu\mathcal{J}_3, \tag{20}$$

$$\mathcal{J}_1 = \sum_{\omega,t} i_{\omega,t} \left(\frac{y_{\omega,t}^{\beta_{\mathrm{NMF}}}}{\beta_{\mathrm{NMF}}\left(\beta_{\mathrm{NMF}}-1\right)} + \frac{z_{\omega,t}^{\beta_{\mathrm{NMF}}}}{\beta_{\mathrm{NMF}}} - \frac{y_{\omega,t} z_{\omega,t}^{\beta_{\mathrm{NMF}}-1}}{\beta_{\mathrm{NMF}}-1} \right), \tag{21}$$

$$\mathcal{J}_2 = \sum_{\omega,t} \overline{i_{\omega,t}} \frac{\left(\sum_k f_{\omega,k} g_{k,t}\right)^{\beta_{\mathrm{reg}}}}{\beta_{\mathrm{reg}}}, \tag{22}$$

$$\mathcal{J}_3 = \sum_{k,l} \left(\sum_{\omega} f_{\omega,k} h_{\omega,l} \right)^2. \tag{23}$$

First, I define the upper bound function for $\mathcal{J}_1$. The second term of $\mathcal{J}_1$ is convex for $\beta_{\mathrm{NMF}} \geq 1$ and concave for $\beta_{\mathrm{NMF}} < 1$, and the third term is convex for $\beta_{\mathrm{NMF}} \geq 2$ and concave for $\beta_{\mathrm{NMF}} < 2$. 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

$$\mathcal{J}_1 \leq \mathcal{J}_1^+ = \sum_{\omega,t} i_{\omega,t} \mathcal{P}_{\omega,t}^{(\beta_{\mathrm{NMF}})}, \tag{24}$$

where

$$\mathcal{P}_{\omega,t}^{(\beta_{\mathrm{NMF}})} = \begin{cases} \mathcal{N}_{\omega,t}^{(\beta_{\mathrm{NMF}})} - y_{\omega,t}\mathcal{M}_{\omega,t}^{(\beta_{\mathrm{NMF}}-1)} & (\beta_{\mathrm{NMF}} < 1) \\ \mathcal{M}_{\omega,t}^{(\beta_{\mathrm{NMF}})} - y_{\omega,t}\mathcal{M}_{\omega,t}^{(\beta_{\mathrm{NMF}}-1)} & (1 \leq \beta_{\mathrm{NMF}} \leq 2) \\ \mathcal{M}_{\omega,t}^{(\beta_{\mathrm{NMF}})} - y_{\omega,t}\mathcal{N}_{\omega,t}^{(\beta_{\mathrm{NMF}}-1)} & (\beta_{\mathrm{NMF}} > 2) \end{cases}, \tag{25}$$

$$\mathcal{M}_{\omega,t}^{(\beta_{\mathrm{NMF}})} = \frac{1}{\beta_{\mathrm{NMF}}} \left\{ \sum_k \alpha_{\omega,t,k}\eta_1 \left(\frac{f_{\omega,k} g_{k,t}}{\alpha_{\omega,t,k}\eta_1} \right)^{\beta_{\mathrm{NMF}}} + \sum_l \gamma_{\omega,t,l}\eta_2 \left(\frac{h_{\omega,l} u_{l,t}}{\gamma_{\omega,t,l}\eta_2} \right)^{\beta_{\mathrm{NMF}}} \right\}, \tag{26}$$

$$\mathcal{N}_{\omega,t}^{(\beta_{\mathrm{NMF}})} = \sigma_{\omega,t}^{\beta_{\mathrm{NMF}}-1} \left(z_{\omega,t} - \sigma_{\omega,t} \right) + \frac{\sigma_{\omega,t}^{\beta_{\mathrm{NMF}}}}{\beta_{\mathrm{NMF}}}. \tag{27}$$

The equality in (24) holds if and only if the auxiliary variables are set as follows:

$$\alpha_{\omega,t,k} = \frac{f_{\omega,k} g_{k,t}}{\sum_{k'} f_{\omega,k'} g_{k',t}}, \tag{28}$$

$$\gamma_{\omega,t,l} = \frac{h_{\omega,l} u_{l,t}}{\sum_{l'} h_{\omega,l'} u_{l',t}}, \tag{29}$$

$$\eta_1 = \frac{\sum_{k'} f_{\omega,k'} g_{k',t}}{\sum_{k'} f_{\omega,k'} g_{k',t} + \sum_{l'} h_{\omega,l'} u_{l',t}}, \tag{30}$$

$$\eta_2 = \frac{\sum_{l'} h_{\omega,l'} u_{l',t}}{\sum_{k'} f_{\omega,k'} g_{k',t} + \sum_{l'} h_{\omega,l'} u_{l',t}}, \tag{31}$$

$$\sigma_{\omega,t} = \sum_{k'} f_{\omega,k'} g_{k',t} + \sum_{l'} h_{\omega,l'} u_{l',t}. \tag{32}$$

Second, I define the upper bound function for $\mathcal{J}_2$. This term is convex for $\beta_{\mathrm{reg}} \geq 1$ and concave for $\beta_{\mathrm{reg}} < 1$. Similarly to (24)-(27), we can define the upper bound function $\mathcal{J}_2^+$ using auxiliary variables $\alpha_{\omega,t,k}$ and $\rho_{\omega,t}$ as

$$\mathcal{J}_2 \leq \mathcal{J}_2^+ = \sum_{\omega,t} \overline{i_{\omega,t}} \mathcal{S}_{\omega,t}^{(\beta_{\mathrm{reg}})}, \tag{33}$$

where

$$\mathcal{S}_{\omega,t}^{(\beta_{\mathrm{reg}})} = \begin{cases} \rho_{\omega,t}^{\beta_{\mathrm{reg}}-1} \left(\sum_k f_{\omega,k} g_{k,t} - \rho_{\omega,t} \right) + \dfrac{\rho_{\omega,t}^{\beta_{\mathrm{reg}}}}{\beta_{\mathrm{reg}}} & (\beta_{\mathrm{reg}} < 1) \\ \dfrac{1}{\beta_{\mathrm{reg}}} \sum_k \alpha_{\omega,t,k} \left(\dfrac{f_{\omega,k} g_{k,t}}{\alpha_{\omega,t,k}} \right)^{\beta_{\mathrm{reg}}} & (1 \leq \beta_{\mathrm{reg}}) \end{cases}. \tag{34}$$

The equality in (33) holds if and only if the auxiliary variables are set as (28) and as follows:

$$\rho_{\omega,t} = \sum_{k'} f_{\omega,k'} g_{k',t}. \tag{35}$$

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

$$\mathcal{J}_3 \leq \mathcal{J}_3^+ = \sum_{k,l,\omega} \frac{f_{\omega,k}^2 h_{\omega,l}^2}{\delta_{k,l,\omega}}. \tag{36}$$

The equality in (36) holds if and only if the auxiliary variables are set as follows:

$$\delta_{k,l,\omega} = \frac{f_{\omega,k} h_{\omega,l}}{\sum_{\omega'} f_{\omega',k} h_{\omega',l}}. \tag{37}$$

Finally, using (24), (33), and (36), we can define the upper bound function $\mathcal{J}^{+}(\Theta, \hat{\Theta})$ as

$$\mathcal{J}^{+}(\Theta, \hat{\Theta}) = \mathcal{J}_1^{+} + \lambda \mathcal{J}_2^{+} + \mu \mathcal{J}_3^{+}, \tag{38}$$

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

$$\sum_{\omega} i_{\omega,t} \left(\mathcal{V}_{\beta_{\mathrm{NMF}}}(\Theta, \hat{\Theta}) - \mathcal{W}_{\beta_{\mathrm{NMF}}}(\Theta, \hat{\Theta}) \right) + \lambda \mathcal{X}_{\beta_{\mathrm{reg}}}(\Theta, \hat{\Theta}) = 0, \tag{39}$$

where

$$\mathcal{V}_{\beta_{\mathrm{NMF}}}(\Theta, \hat{\Theta}) = \begin{cases} \sigma_{\omega,t}^{\beta_{\mathrm{NMF}}-1} f_{\omega,k} & (\beta_{\mathrm{NMF}} < 1) \\ g_{k,t}^{\beta_{\mathrm{NMF}}-1} \left(\alpha_{k,\omega,t} \eta_1 \right)^{1-\beta_{\mathrm{NMF}}} f_{\omega,k}^{\beta_{\mathrm{NMF}}} & (1 \leq \beta_{\mathrm{NMF}}) \end{cases}, \tag{40}$$

$$\mathcal{W}_{\beta_{\mathrm{NMF}}}(\Theta, \hat{\Theta}) = \begin{cases} y_{\omega,t} g_{k,t}^{\beta-2} \left(\alpha_{k,\omega,t} \eta_1 \right)^{2-\beta_{\mathrm{NMF}}} f_{\omega,k}^{\beta_{\mathrm{NMF}}-1} & (\beta_{\mathrm{NMF}} \leq 2) \\ y_{\omega,t} \sigma_{\omega,t}^{\beta_{\mathrm{NMF}}-2} f_{\omega,k} & (2 < \beta_{\mathrm{NMF}}) \end{cases}, \tag{41}$$

$$\mathcal{X}_{\beta_{\mathrm{reg}}}(\Theta, \hat{\Theta}) = \begin{cases} \sum_{\omega} \overline{i_{\omega,t}} \rho_{\omega,t}^{\beta_{\mathrm{reg}}-1} f_{\omega,k} & (\beta_{\mathrm{reg}} < 1) \\ \sum_{\omega} \overline{i_{\omega,t}} f_{\omega,k} \left(\dfrac{f_{\omega,k} g_{k,t}}{\alpha_{\omega,t,k}} \right)^{\beta_{\mathrm{reg}}-1} & (1 \leq \beta_{\mathrm{reg}}) \end{cases}. \tag{42}$$

By solving (39) for $g_{k,t}$ under the nonnegativity, we obtain

$$g_{k,t} = \begin{cases} \left(\dfrac{\sum_{\omega} i_{\omega,t} y_{\omega,t} \left(\alpha_{k,\omega,t} \eta_1 \right)^{2-\beta_{\mathrm{NMF}}} f_{\omega,k}^{\beta_{\mathrm{NMF}}-1}}{\sum_{\omega} i_{\omega,t} \sigma_{\omega,t}^{\beta_{\mathrm{NMF}}-1} f_{\omega,k} + \lambda \mathcal{X}_{\beta_{\mathrm{reg}}}} \right)^{\frac{1}{2-\beta_{\mathrm{NMF}}}} & (\beta_{\mathrm{NMF}} < 1) \\ \dfrac{\sum_{\omega} i_{\omega,t} y_{\omega,t} \left(\alpha_{k,\omega,t} \eta_1 \right)^{2-\beta_{\mathrm{NMF}}} f_{\omega,k}^{\beta_{\mathrm{NMF}}-1} g_{k,t}^{\beta_{\mathrm{NMF}}-1}}{\sum_{\omega} i_{\omega,t} g_{k,t}^{\beta_{\mathrm{NMF}}-1} \left(\alpha_{k,\omega,t} \eta_1 \right)^{1-\beta_{\mathrm{NMF}}} f_{\omega,k}^{\beta_{\mathrm{NMF}}} + \lambda \mathcal{X}_{\beta_{\mathrm{reg}}}} & (1 \leq \beta_{\mathrm{NMF}} \leq 2) \\ \left(\dfrac{\sum_{\omega} i_{\omega,t} y_{\omega,t} \sigma_{\omega,t}^{\beta_{\mathrm{NMF}}-2} f_{\omega,k} g_{k,t}^{\beta_{\mathrm{NMF}}-1}}{\sum_{\omega} i_{\omega,t} g_{k,t}^{\beta_{\mathrm{NMF}}-1} \left(\alpha_{k,\omega,t} \eta_1 \right)^{1-\beta_{\mathrm{NMF}}} f_{\omega,k}^{\beta_{\mathrm{NMF}}} + \lambda \mathcal{X}_{\beta_{\mathrm{reg}}}} \right)^{\frac{1}{\beta_{\mathrm{NMF}}-1}} & (2 < \beta_{\mathrm{NMF}}) \end{cases}. \tag{43}$$

Then we can obtain the update rule of $g_{k,t}$ by substituting (28), (30), (32), and (35) into (43) as follows:

$$g_{k,t} \leftarrow g_{k,t} \left(\frac{\sum_\omega i_{\omega,t} y_{\omega,t} f_{\omega,k} z_{\omega,t}^{\beta_{\mathrm{NMF}}-2}}{\sum_\omega i_{\omega,t} f_{\omega,k} z_{\omega,t}^{\beta_{\mathrm{NMF}}-1} + \lambda \sum_\omega \overline{i_{\omega,t}} f_{\omega,k} \left(\sum_{k'} f_{\omega,k'} g_{k',t}\right)^{\beta_{\mathrm{reg}}-1}} \right)^{\varphi(\beta_{\mathrm{NMF}})}. \tag{44}$$

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

$$h_{\omega,l} \leftarrow h_{\omega,l} \left(\frac{\sum_t i_{\omega,t} y_{\omega,t} u_{l,t} z_{\omega,t}^{\beta_{\mathrm{NMF}}-2}}{\sum_t i_{\omega,t} u_{l,t} z_{\omega,t}^{\beta_{\mathrm{NMF}}-1} + 2\mu \sum_k f_{\omega,k} \sum_{\omega'} f_{\omega',k} h_{\omega',l}} \right)^{\varphi(\beta_{\mathrm{NMF}})}, \tag{45}$$

$$u_{l,t} \leftarrow u_{l,t} \left(\frac{\sum_\omega i_{\omega,t} y_{\omega,t} h_{\omega,l} z_{\omega,t}^{\beta_{\mathrm{NMF}}-2}}{\sum_\omega i_{\omega,t} h_{\omega,l} z_{\omega,t}^{\beta_{\mathrm{NMF}}-1}} \right)^{\varphi(\beta_{\mathrm{NMF}})}. \tag{46}$$

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

3.3 Theoretical analysis of basis extrapolation based on generation model

3.3.1 Optimal divergence for basis extrapolation and generation model

The proposed method attempts both signal *separation* and basis *extrapolation* using the supervised bases $\boldsymbol{F}$. In previous studies, the analysis of optimal divergence only for signal separation has been discussed [21, 22, 31]. However, there has been no discussion on the optimal divergence for the extrapolation techniques using NMF. In this section, I analyze the extrapolation ability based on a statistical generation model of the observed data $\boldsymbol{Y}$, and determine the optimal divergence for basis extrapolation w.r.t. various β_{NMF} and β_{reg} values.

In NMF decomposition, the minimization of β-divergence between $\boldsymbol{Y}$ and $\boldsymbol{FG}$ corresponds to a log-likelihood maximization under the assumption of the generation model of $\boldsymbol{Y}$ for each β_{NMF} [34]. The minimization of $\mathcal{D}_{\beta_{\mathrm{NMF}}}(y_{\omega,t} \| \vartheta)$ is equivalent to the maximization of $\exp(-\mathcal{D}_{\beta_{\mathrm{NMF}}}(y_{\omega,t} \| \vartheta))$. Here, we can rewrite

$\exp(-\mathcal{D}_{\beta_{\mathrm{NMF}}}(y_{\omega,t}\|\vartheta))$ as

$$\exp\left(-\mathcal{D}_{\beta_{\mathrm{NMF}}}(y_{\omega,t}\|\vartheta)\right) = \begin{cases} \dfrac{y_{\omega,t}}{\vartheta}\exp\left(-\dfrac{y_{\omega,t}}{\vartheta}+1\right) & (\beta_{\mathrm{NMF}}=0) \\ \left(\dfrac{\vartheta e}{y_{\omega,t}}\right)^{y_{\omega,t}}\exp\left(-\vartheta\right) & (\beta_{\mathrm{NMF}}=1) \\ \exp\left(-\dfrac{\left(y_{\omega,t}-\vartheta\right)^2}{2}\right) & (\beta_{\mathrm{NMF}}=2) \\ \exp\left(\dfrac{\vartheta^{\beta_{\mathrm{NMF}}-1}y_{\omega,t}}{\beta_{\mathrm{NMF}}-1}-\dfrac{\vartheta^{\beta_{\mathrm{NMF}}}}{\beta_{\mathrm{NMF}}}\right) & (\beta_{\mathrm{NMF}}\geq 3) \end{cases}, \tag{47}$$

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 (47) is given by

$$y_{\omega,t} \sim p\left(y_{\omega,t}\right) = \begin{cases} \dfrac{1}{\vartheta_1}\exp\left(-\dfrac{y_{\omega,t}}{\vartheta_1}\right) & (\beta_{\mathrm{NMF}}=0) \\ \dfrac{{\vartheta_2}^{y_{\omega,t}}}{\Gamma\left(y_{\omega,t}+1\right)}\exp\left(-\vartheta_2\right) & (\beta_{\mathrm{NMF}}=1) \\ \dfrac{1}{\sqrt{2\pi\vartheta_3}}\exp\left(-\dfrac{\left(y_{\omega,t}-\vartheta_4\right)^2}{2{\vartheta_3}^2}\right) & (\beta_{\mathrm{NMF}}=2) \\ C\exp\left(\dfrac{{\vartheta_5}^{\beta_{\mathrm{NMF}}-1}y_{\omega,t}}{\beta_{\mathrm{NMF}}-1}\right) & (\beta_{\mathrm{NMF}}\geq 3) \end{cases}, \tag{48}$$

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

$$C = \vartheta_5^{\beta_{\mathrm{NMF}}-1}\left(\beta_{\mathrm{NMF}}-1\right)^{-1}\left(\exp\left(\frac{\vartheta_5^{\beta_{\mathrm{NMF}}-1}}{\beta_{\mathrm{NMF}}-1}C_m\right)-1\right)^{-1}. \tag{49}$$

Using (48), we can generate the most probable spectrogram for each β_{NMF}.

3.3.2 Simulation conditions

To analyze the net extrapolation ability, I simulate the spectrogram restoration task. In this simulation, I generated random i.i.d. values, which obey the corresponding generation model (48) for each β_{NMF}, as the observed data matrix $\boldsymbol{Y}$. I compared $\beta_{\mathrm{NMF}} = 0, 1, 2, 3, 4$ and $\beta_{\mathrm{reg}} = 0, 1, 2, 3$, and I used the same divergence β_{NMF} in the training and separation processes. The size of this data matrix was set to $\Omega = 5000$ and $T = 200$. I 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 are determined so as to generate the nonnegative random i.i.d. values that obey each corresponding generation model. Note that the parameters θ_1–θ_5 simply determine the scales of the input random variables, and basically can be set to arbitrary value without loss of generality. In addition, I used two types of data-missing patterns $\boldsymbol{I}$, in which 75% or 98% of the grids were missing in a uniform manner, and the missing data $\boldsymbol{I} \circ \boldsymbol{Y}$ imitated the binary-masking procedure. The supervised bases $\boldsymbol{F}$ were obtained by training using the same data matrix $\boldsymbol{Y}$, namely, $\boldsymbol{Y}_{\mathrm{target}} = \boldsymbol{Y}$ in Fig. 9. The number of supervised bases, K, was 100, which is the half size of T, and the number of other bases, L, was 30. Therefore, the task was to reconstruct original $\boldsymbol{Y}$ from the observations with missing data, $\boldsymbol{I} \circ \boldsymbol{Y}$, using the trained bases.

3.3.3 Simulation results and discussion

I used sources-to-artifacts ratio (SAR) defined in [35] as the accuracy of the extrapolation. Here, the estimated signal $\hat{s}(t)$ is defined as

$$\hat{s}(t) = s_{\mathrm{target}}(t) + s_{\mathrm{interf}}(t) + s_{\mathrm{artif}}(t), \tag{50}$$

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

$$\mathrm{SAR} = 10 \log_{10} \frac{\sum_t \left\{ s_{\mathrm{target}}(t) + e_{\mathrm{interf}}(t) \right\}^2}{\sum_t e_{\mathrm{artif}}(t)^2}. \tag{51}$$

Therefore, SAR indicates the absence of artificial distortion.

Figure 15 shows the SAR result for each divergence and regularization. From this result, it is confirmed that a higher β_{NMF} provides better basis extrapolation regardless of the type of regularization (β_{reg}). In NMF decomposition, if we set β_{NMF} to a large value, the trained bases tend to become anti-sparse (smooth). In contrast, if β_{NMF} is close to zero, the trained bases become more sparsity-aware, and this property is suitable for normal NMF-based music source separation because of the inherent sparsity of music spectrograms (e.g., $\beta_{\mathrm{NMF}} = 1$ is recommended in [21, 22, 31]). 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 $\beta_{\mathrm{NMF}} > 1$ rather than KL-divergence ($\beta_{\mathrm{NMF}} = 1$) because of the trade-off between separation and extrapolation abilities, as illustrated in Fig. 16. This issue will be confirmed experimentally in the next section.

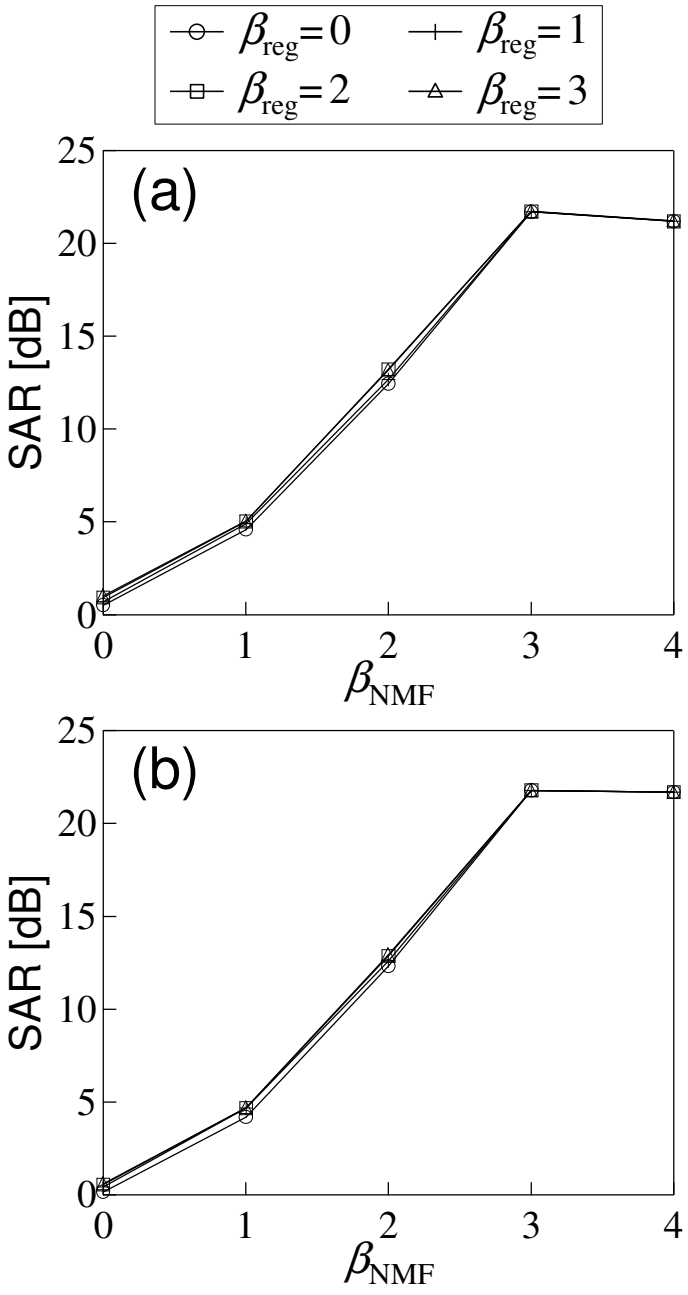

Figure 15. Extrapolation abilities for (a) 75%-binary-masked data and (b) 98%-binary-masked data.

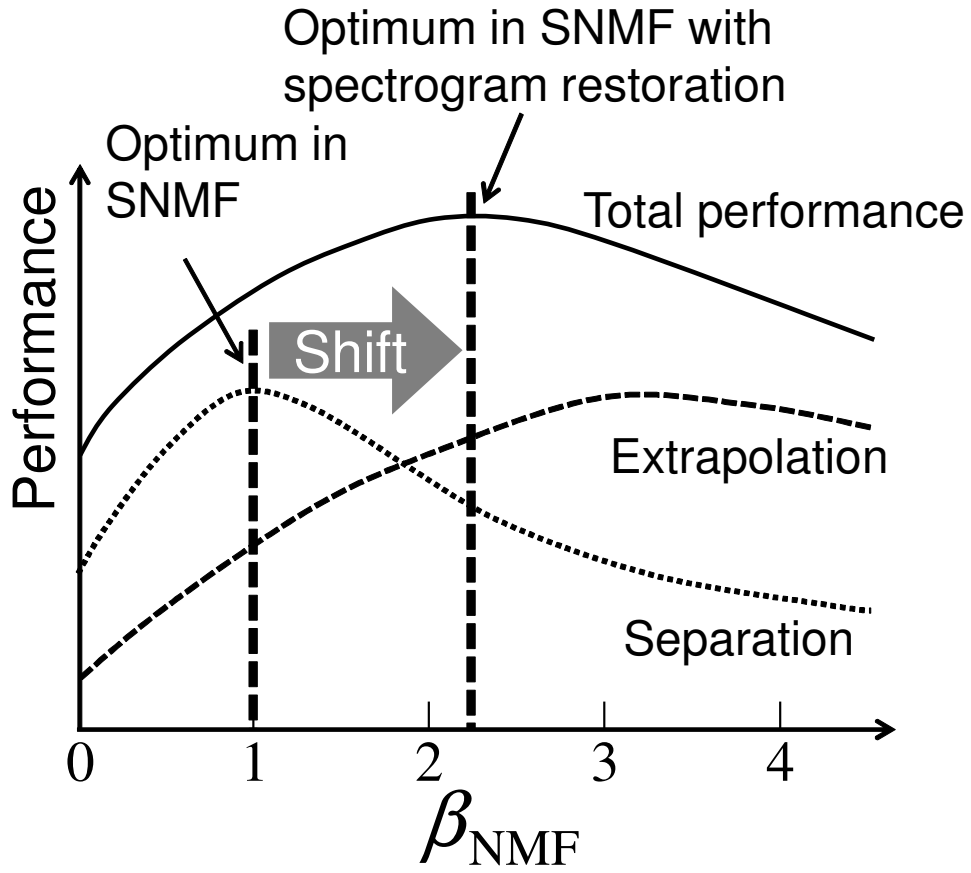

Figure 16. Trade-off between separation and extrapolation abilities. Overall performance is highest when $\beta_{\mathrm{NMF}} > 1$.

3.4 Comparison between proposed hybrid method and conventional methods

3.4.1 Experimental conditions

I conducted objective evaluation to confirm the effectiveness of the proposed hybrid method described in the previous section. In this experiment, I compared five methods, namely, simple directional clustering [26], Multichannel NMF based on IS-divergence [25], PSNMF [21, 22], conventional hybrid method that concatenates PSNMF after directional clustering [29], and proposed hybrid method including SNMF with spectrogram restoration after directional clustering, in terms

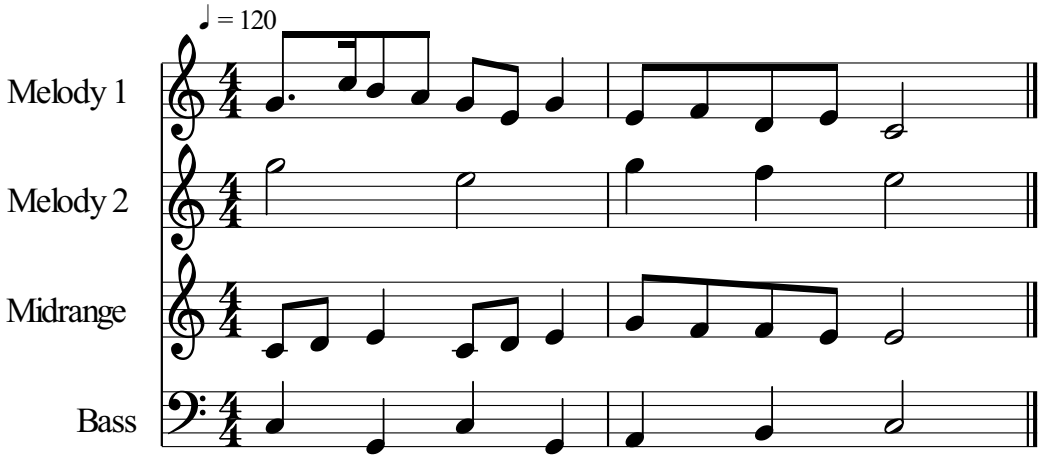

Figure 17. Scores of each part.

Table 1. 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

of their ability to separate music artificial and real-recorded signals. Also, I compared some evaluation scores with various β_{NMF} and β_{reg} for PSNMF and the proposed hybrid method by setting five divergences and regularizations, namely, $\beta = 0, 1, 2, 3$, and 4. I used the same divergence (β_{NMF}) in the training and separation processes for PSNMF and proposed SNMF with spectrogram restoration in the proposed hybrid method. In this experiment, I conducted two experiments to consider artificial signal and real-recorded signal cases. I used stereo signals containing four melody parts (depicted in Fig. 17) with three compositions (C1–C3) of instruments shown in Table 1. These signals were artificially generated by a MIDI synthesizer. In particular, these stereo signals were mixed down to a monaural format only when we evaluate the separation accuracy of PSNMF because PSNMF is a separation method for a monaural input signal.

In the artificial signal case, the observed signals $\boldsymbol{Y}$ were produced by mixing

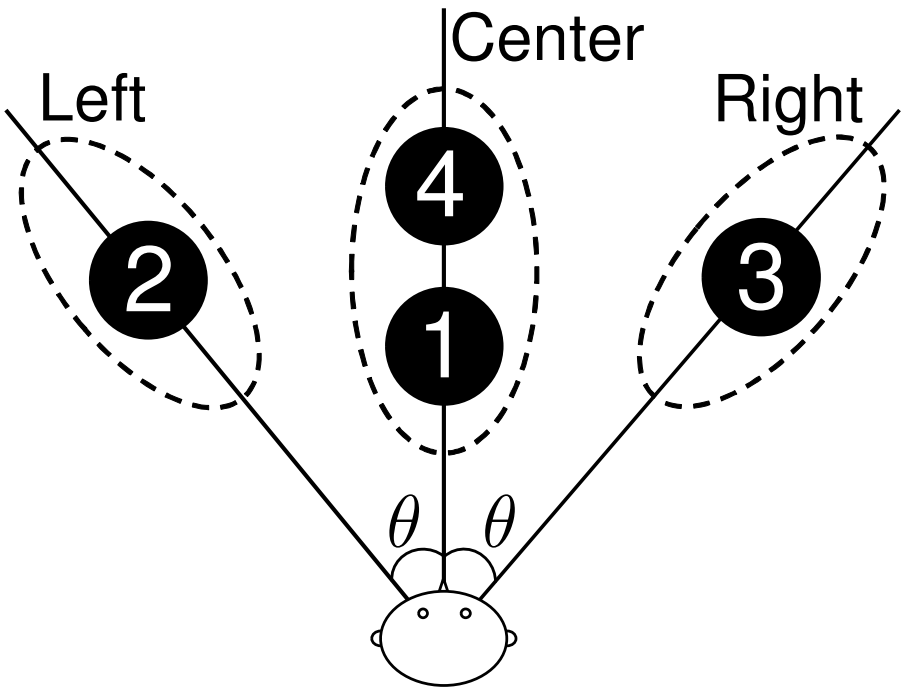

Figure 18. Panning of four sources with sine law used in artificial signal case experiment. Numbered black circles represent locations of instruments in stereo format. For example, if target is Ob., No.1 is set to Ob. and Nos.2, 3, and 4 are combinations of Fl., Tb., and Pf.

four sources with the same power. The observed signal contained one source in the left and right directions and two sources in the center direction based on a sine law (see Fig. 18). The target instrument is always located in the center direction along with another interfering instrument, and we prepared two patterns in which the left and right sources are located at $\theta = 15^\circ$ and 45°. In addition, I 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 and separation were 500. Moreover, the number of clusters used in directional clustering was 3, the number of a priori bases, K, was 100, and the number of bases for matrix $\boldsymbol{H}$, H, was 30. The weighting parameters λ and μ were empirically determined.

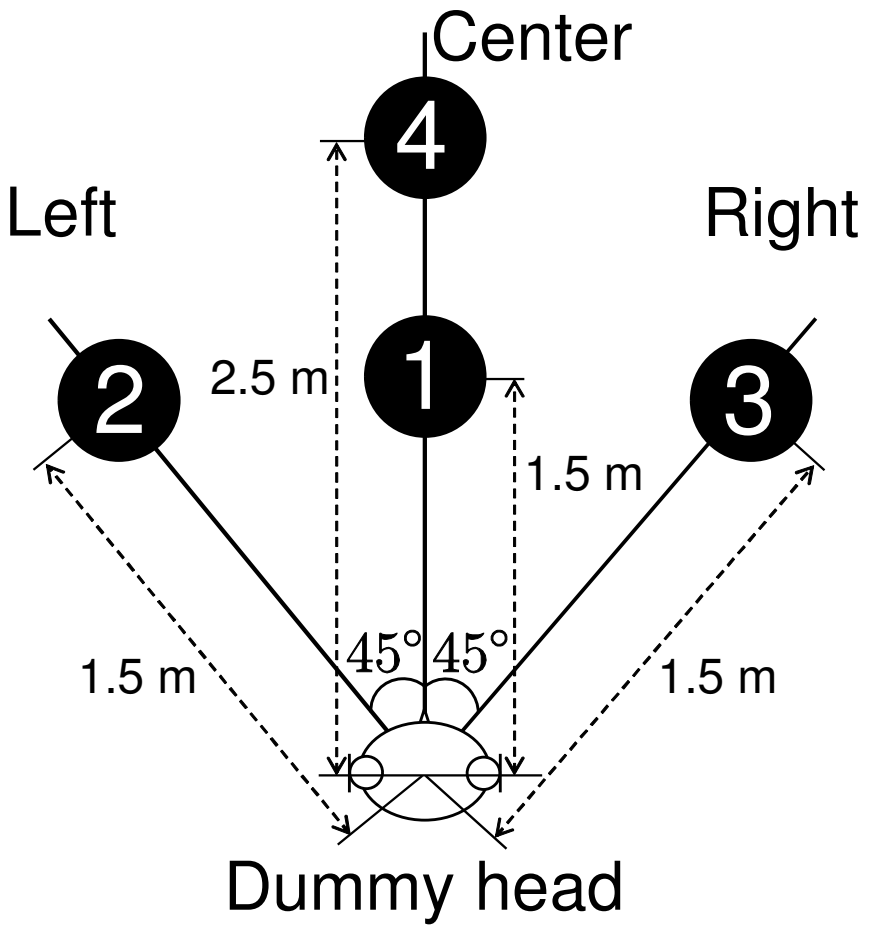

Figure 19. Geometry of the loudspeaker and binaural microphone (dummy head). Numbered black circles represent locations of loudspeakers. Target source and supervision sound is always located in No.1 position.

In the real-recorded signal case, I recorded each instrumental solo signal and the supervision sound, which are the same as those in the artificial case, in an experimental room whose reverberation time was 200 ms. A geometry of the loudspeaker and binaural microphone NEUMANN KU-100 is shown in Fig. 19. The target source and the supervision sound is always located in No.1 position in Fig. 19. The observed signal $\boldsymbol{Y}$ was produced by mixing these recorded signals as the same power. Other conditions were the same as those of the artificial signal case.

3.4.2 Experimental results

I used the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), and SAR defined in [35] as the evaluation scores. The formulas for SDR and SIR are defined as

$$\mathrm{SDR} = 10\log_{10}\frac{\sum_t s_{\mathrm{target}}(t)^2}{\sum_t \{e_{\mathrm{interf}}(t) + e_{\mathrm{artif}}(t)\}^2}, \tag{52}$$

$$\mathrm{SIR} = 10\log_{10}\frac{\sum_t s_{\mathrm{target}}(t)^2}{\sum_t e_{\mathrm{interf}}(t)^2}. \tag{53}$$

SDR indicates the quality of the separated target sound, and SIR indicates the degree of separation between the target and other sounds. Therefore, SDR indicates the total evaluation score that involves SIR and SAR.

First, I compare the variation of separation performance for various β_{NMF} and β_{reg}. Figures 20 and 21 show the average SDR, SIR, and SAR of the proposed hybrid method for each divergence (β_{NMF}) and each regularization (β_{reg}) in the artificial signal case with $\theta = 15^\circ$ and $\theta = 45^\circ$, where the four instruments are shuffled with 12 combinations in each of compositions C1–C3. Therefore, these results are the averages of 36 input signal patterns. Also, Fig. 22 show the average SDR, SIR, and SAR in the real-recorded signal case. From the SDRs in Figs. 20, 21, and 22, the regularization with KL-divergence ($\beta_{\mathrm{reg}} = 1$) is slightly better than the other divergences but the difference is not significant, except for the case of $\beta_{\mathrm{reg}} = 0$. In addition, we can confirm that the EUC-distance-based cost function ($\beta_{\mathrm{NMF}} = 2$) is an optimal divergence for the proposed hybrid method including SNMF with spectrogram restoration.

Next, I compare the separation performance of the proposed hybrid method with the other conventional methods, where I compare the evaluation scores of the proposed hybrid method only when $\beta_{\mathrm{reg}} = 1$ because this KL-divergence-based regularization achieves the highest separation performance. Figures 23 and 24 show the average SDR, SIR, and SAR for each method in the artificial signal case with $\theta = 15^\circ$ and $\theta = 45^\circ$. Also, Fig. 25 shows the average SDR, SIR, and SAR in the real-recorded signal case. Similarly to the previous results, these results are the averages of 36 input signal patterns, which contain all compositions and instrumental combinations.

From the SDRs in Figs. 23, 24, and 25, we can confirm that directional clustering does not have sufficient performance because this method cannot discriminate the sources in the same direction. Multichannel NMF also cannot achieve the sufficient separation because this method strongly depends on the initial value and lack robustness. In contrast, the methods using SNMF can give better results and the proposed hybrid method with SNMF with spectrogram restoration outperforms all other methods in both artificial and real-recorded signal cases. In addition, the conventional hybrid method is inferior to PSNMF when $\beta_{\mathrm{NMF}} \leq 1$ whereas this hybrid method utilizes both directional clustering and PSNMF. This is because the conventional hybrid method is affected by the spectral chasms and cannot reconstruct such lost components. Furthermore, we can confirm that the EUC-distance-based cost function ($\beta_{\mathrm{NMF}} = 2$) is an optimal divergence for the proposed hybrid method, whereas KL-divergence ($\beta_{\mathrm{NMF}} = 1$) is the best divergence even for conventional PSNMF [21, 22, 31]. This marked shift of the optimal divergence in SNMF with spectrogram restoration is due to the trade-off between the separation and extrapolation abilities, as predicted in Sect. 3.3.

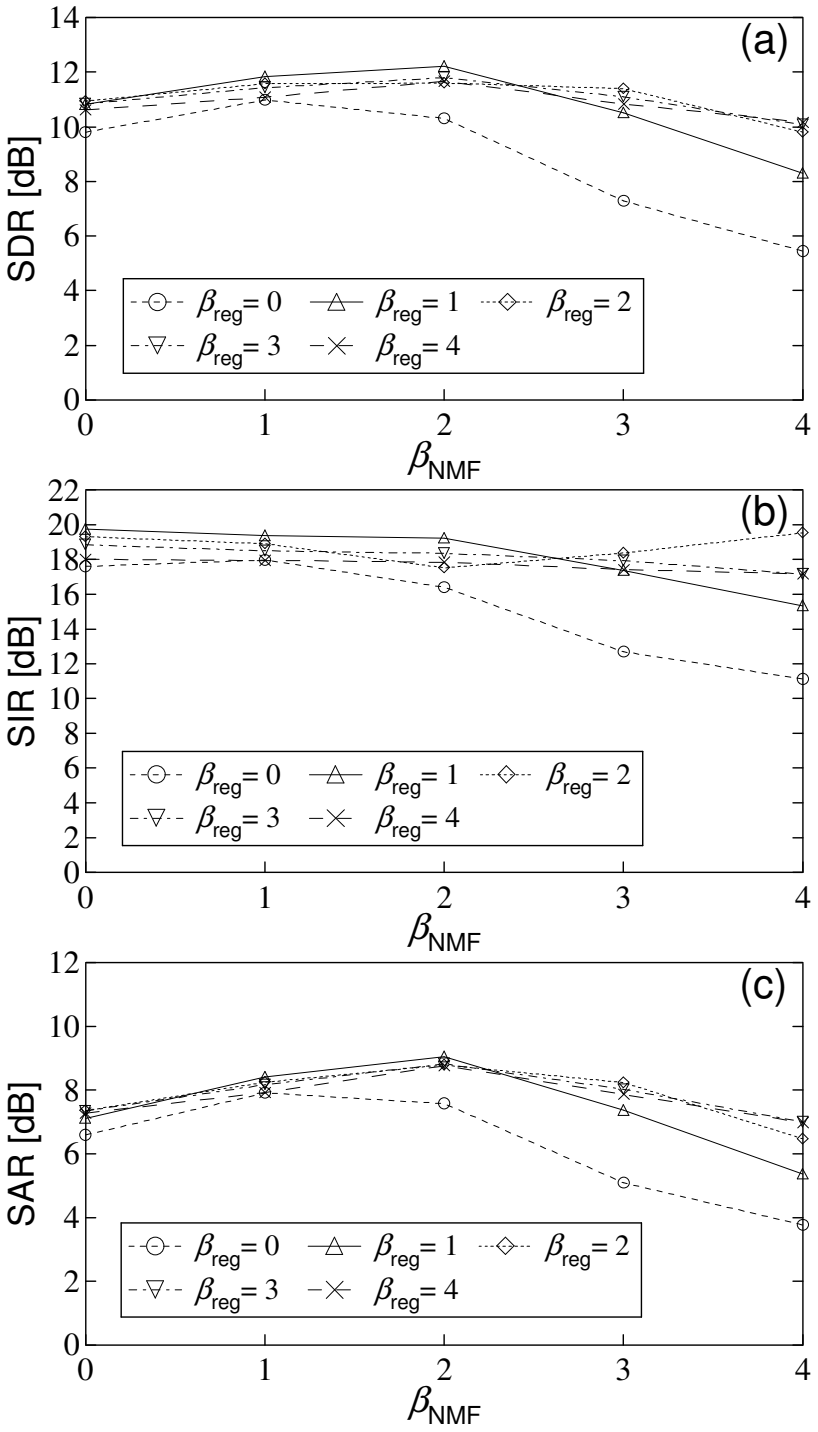

Figure 20. Average scores with various divergences and regularizations in artificial signal case when $\theta = 15°$: (a) shows SDR, (b) shows SIR, and (c) shows SAR for proposed methods.

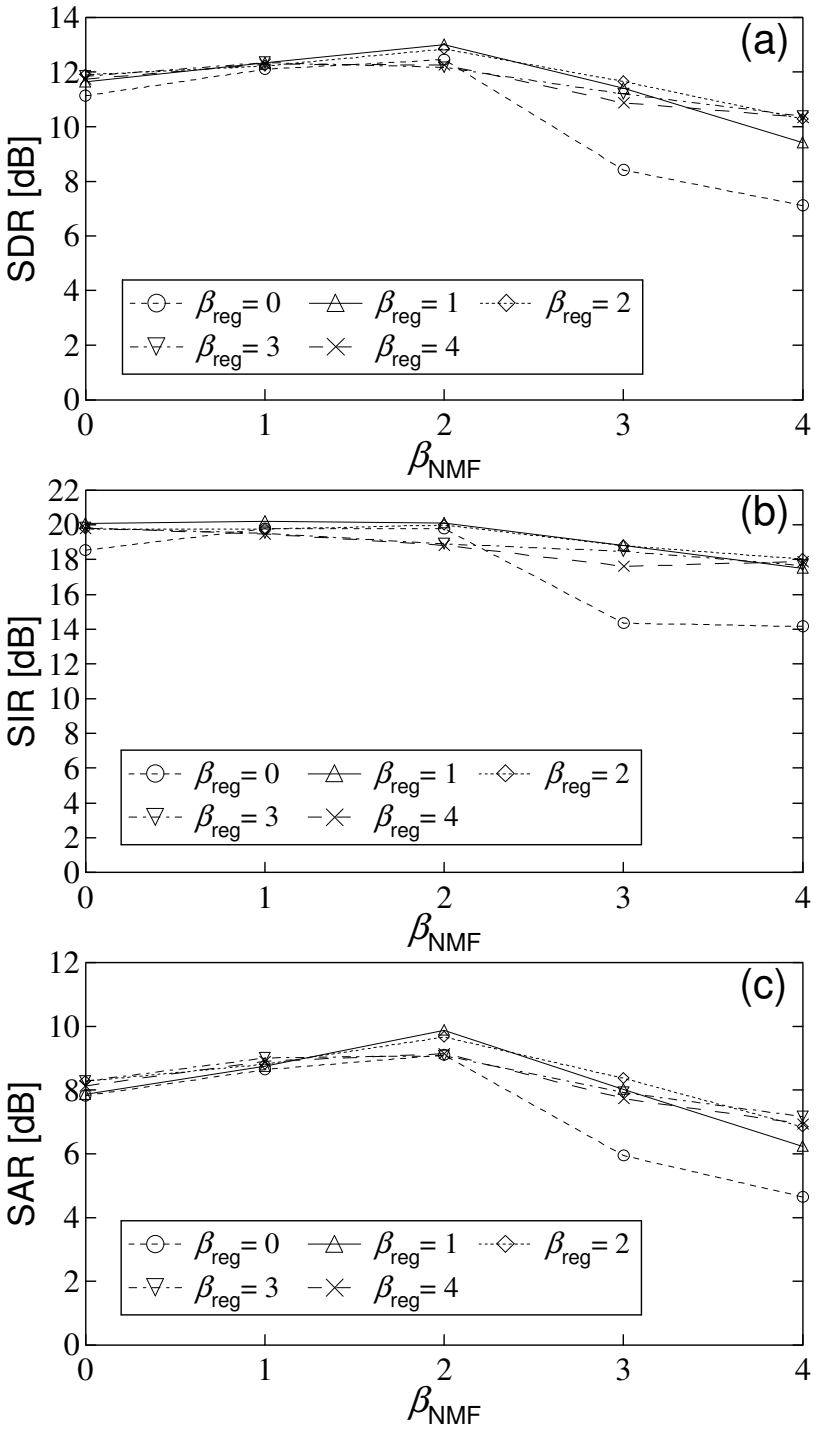

Figure 21. Average scores with various divergences and regularizations in artificial signal case when $\theta = 45°$: (a) shows SDR, (b) shows SIR, and (c) shows SAR for conventional and proposed methods.

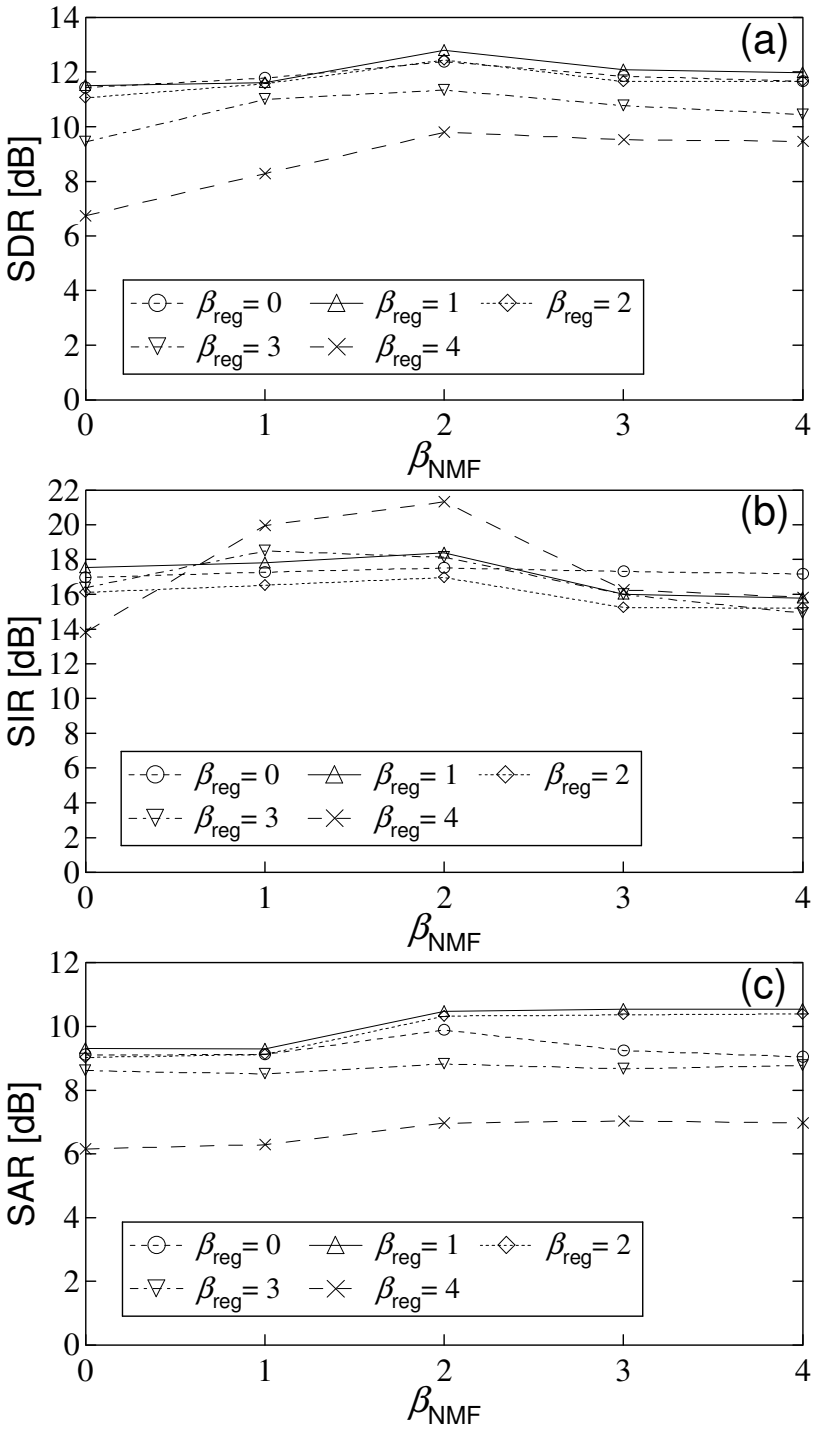

Figure 22. Average scores with various divergences and regularizations in real-recorded signal case: (a) shows SDR, (b) shows SIR, and (c) shows SAR for conventional and proposed methods.

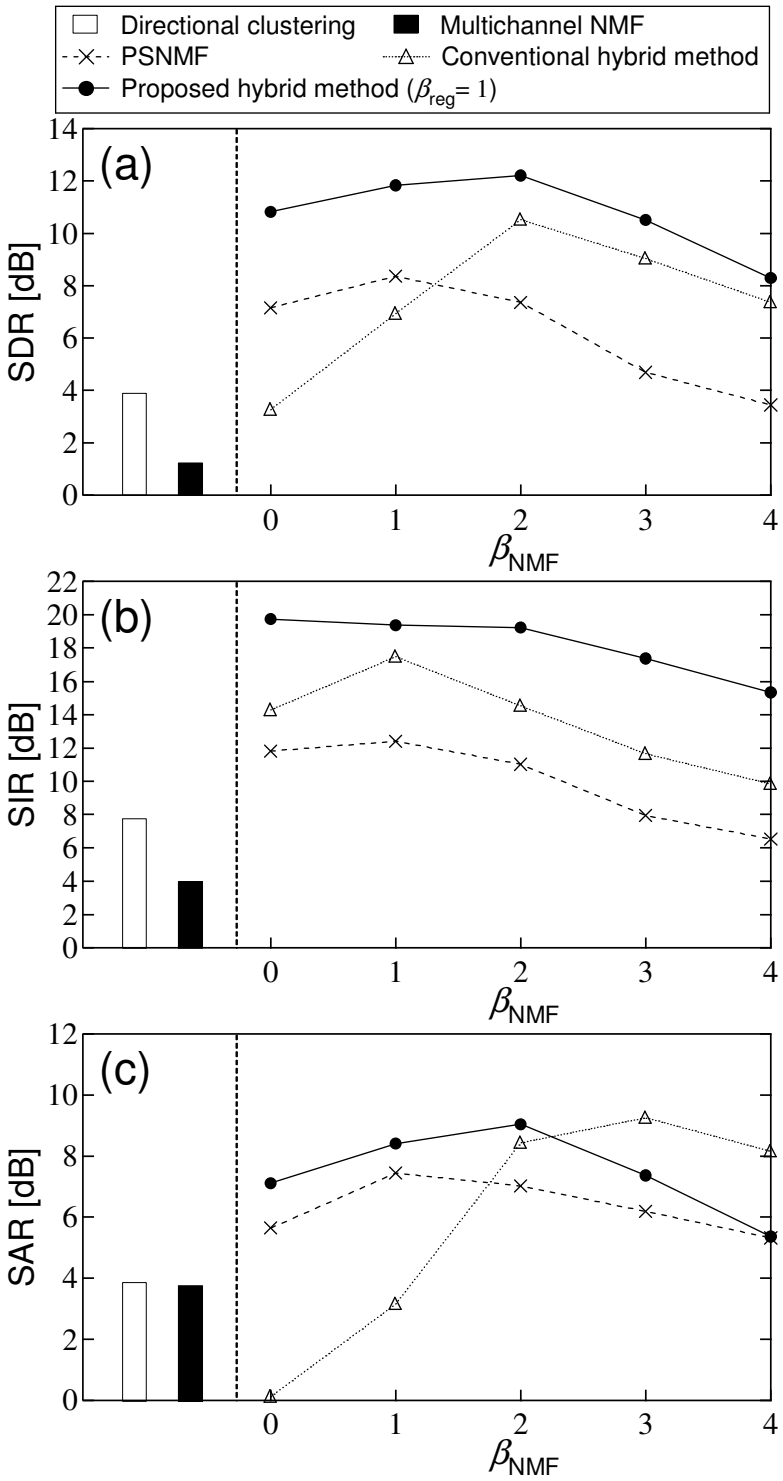

Figure 23. Average scores in artificial signal case when $\theta = 15°$: (a) shows SDR, (b) shows SIR, and (c) shows SAR for conventional and proposed methods.

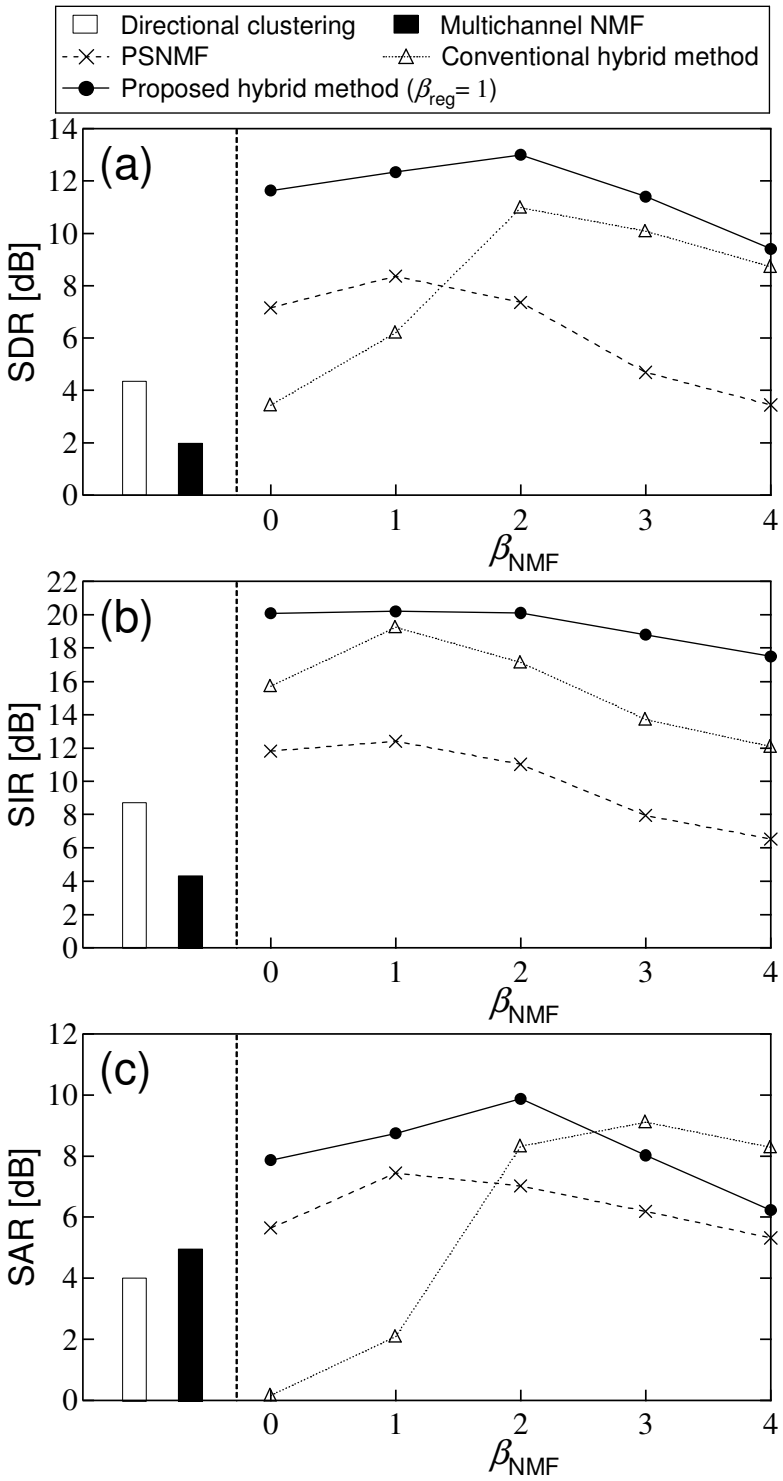

Figure 24. Average scores in artificial signal case when $\theta = 45°$: (a) shows SDR, (b) shows SIR, and (c) shows SAR for conventional and proposed methods.

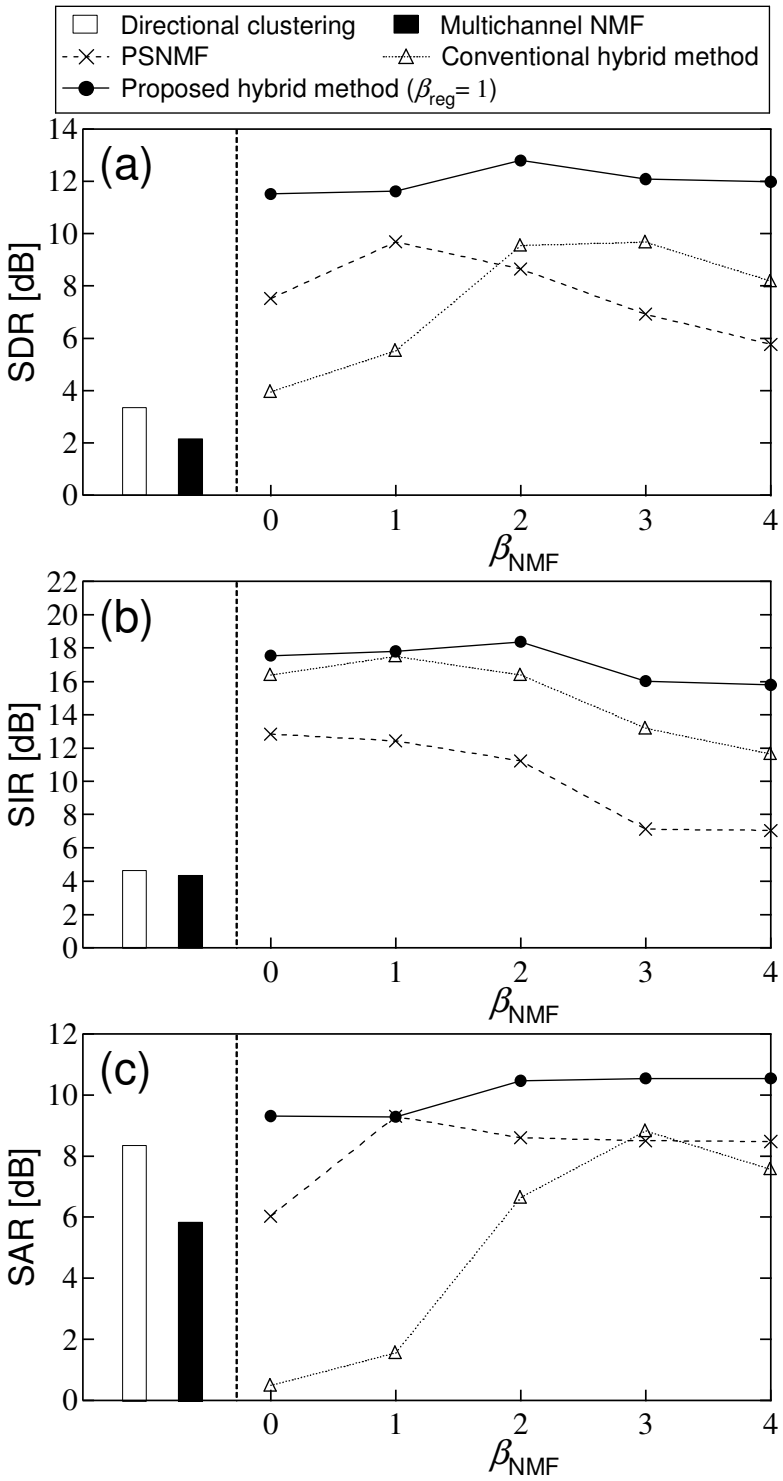

Figure 25. Average scores in real-recorded signal case: (a) shows SDR, (b) shows SIR, and (c) shows SAR for conventional and proposed methods.

3.5 Conclusion

In this section, first, I derived the update rules of SNMF with spectrogram restoration and proposed hybrid method that concatenates SNMF with spectrogram restoration after directional clustering. This SNMF attempts both signal separation and basis extrapolation simultaneously.

Next, I analyzed the net extrapolation ability based on generation models of each divergence. This analysis revealed the mechanism of marked shift of optimal divergence in SNMF with spectrogram restoration and trade-off between separation and extrapolation abilities owing to the difference of sparseness in each divergence.

Finally, the effectiveness of the proposed hybrid method was confirmed by the experiments for artificial and real-recorded signals. The results showed the marked shift of the optimal divergence for the proposed hybrid method because of the trade-off between separation and extrapolation abilities.

4. Optimal Divergence Diversity for SNMF with spectrogram restoration

4.1 Introduction

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

First, I describe about a new scheme of optimal divergence diversity and derive update rules of the proposed method in Sect. 4.2. Second, I conduct an experiment to compare separation performance of the proposed method with divergence diversity and the divergence-fixed hybrid method in Sect. 4.3. Finally, Sect. 4.4 concludes this section.

4.2 SNMF with spectrogram restoration based on multi-divergence

4.2.1 Divergence dependency on local chasms condition

The optimal divergence for SNMF with spectrogram restoration depends on the rate of spectral chasms in each time frame of the spectrogram obtained by preceding directional clustering because of the trade-off between separation and extrapolation abilities. If there are many chasms in a frame of the binary-masked spectrogram, SNMF is preferred to have high extrapolation ability. In contrast, if the rate of chasms is low value, the separation ability is required rather than the extrapolation. Therefore, it is expected that EUC-distance should be used in the frames that have many chasms, and KL-divergence should be used in the

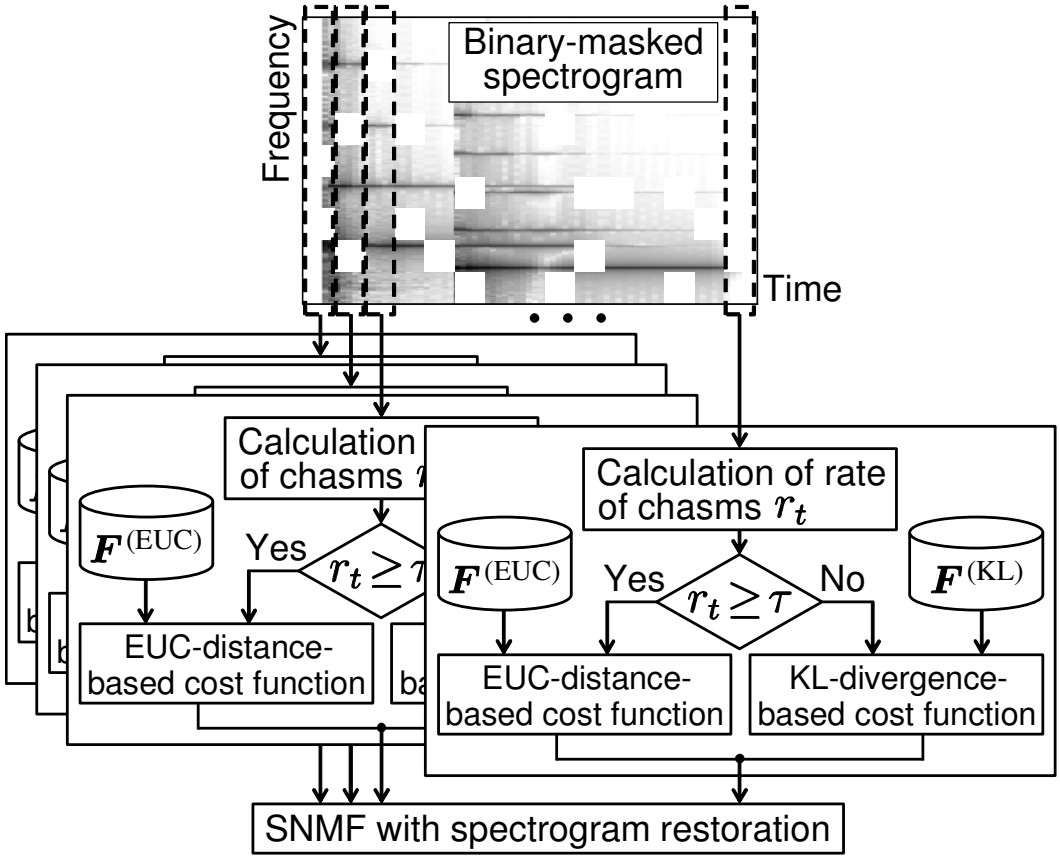

Figure 26. Divergence diversity algorithm of proposed method.

other frames. To improve total separation performance of SNMF with spectrogram restoration for any types of input signals, we propose a new frame-wise divergence switching method as described in the next section.

4.2.2 Cost function and update rules

Considering the above-mentioned divergence dependency on the local chasm condition, we propose to switch the divergence in each frame of the spectrogram according to the rate of chasms in each frame, r_t, and a threshold value τ $(0 \leq \tau \leq 1)$, where the rate of chasms r_t can be calculated from the index matrix $\boldsymbol{I}$. Figure 26 depicts an algorithm of the frame-wise divergence switching. This divergence switching method is implemented by switching the cost function in each frame,

as

$$\mathcal{J}_{\text{diversity}} = \sum_t \mathcal{J}_t, \tag{54}$$

$$\mathcal{J}_t = \begin{cases} \sum_\omega i_{\omega,t} \mathcal{D}_{\beta=2}(y_{\omega,t} \| s^{(\text{EUC})}_{\omega,t}) \\ \quad + \lambda^{(\text{EUC})} \sum_\omega \overline{i_{\omega,t}} \mathcal{D}_{\beta_{\text{reg}}}(0 \| \sum_k f^{(\text{EUC})}_{\omega,k} g_{k,t}) \\ \quad + \mu^{(\text{EUC})} \| \boldsymbol{F}^{(\text{EUC})\text{T}} \boldsymbol{H} \|^2_{\text{Fr}} & (r_t \geq \tau) \\ \sum_\omega i_{\omega,t} \mathcal{D}_{\beta=1}(y_{\omega,t} \| s^{(\text{KL})}_{\omega,t}) \\ \quad + \lambda^{(\text{KL})} \sum_\omega \overline{i_{\omega,t}} \mathcal{D}_{\beta_{\text{reg}}}(0 \| \sum_k f^{(\text{KL})}_{\omega,k} g_{k,t}) \\ \quad + \mu^{(\text{KL})} \| \boldsymbol{F}^{(\text{KL})\text{T}} \boldsymbol{H} \|^2_{\text{Fr}} & (r_t < \tau) \end{cases}, \tag{55}$$

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

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

where $\boldsymbol{F}^{(\text{KL})} (\in \mathbb{R}^{\Omega \times K}_{\geq 0})$ and $\boldsymbol{F}^{(\text{EUC})} (\in \mathbb{R}^{\Omega \times K}_{\geq 0})$ are the supervised basis matrices trained in advance using KL-divergence-based NMF and EUC-distance-based NMF, respectively. Also, $f^{(\text{KL})}_{\omega,k}$ and $f^{(\text{EUC})}_{\omega,k}$ are the entries of $\boldsymbol{F}^{(\text{KL})}$ and $\boldsymbol{F}^{(\text{EUC})}$, respectively, $\mu^{(*)}$ and $\lambda^{(*)}$ are the weighting parameters for each term, and $* = \{\text{KL}, \text{EUC}\}$. The divergence is determined depending on r_t and τ in each frame. Therefore, this method can be considered as a frame-wise *diversity* of the divergence to achieve both of optimal separation and extrapolation.

The update rules based on (54) is obtained by the auxiliary function approach. Similarly to Sect. 3.2.2, we can design the upper bound function $\mathcal{J}^+$ 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

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

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

where

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

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

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

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

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

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

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

The equality in (59) holds if and only if the auxiliary variables are set as (29)

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 from the derivative of the upper bound function (58) w.r.t. each objective variable and substitution of the equality condition (67)–(71), as

$$g_{k,t} \leftarrow \begin{cases} g_{k,t} \cdot \dfrac{\sum_\omega i_{\omega,t} y_{\omega,t} f^{(\mathrm{EUC})}_{\omega,k}}{\sum_\omega i_{\omega,t} f^{(\mathrm{EUC})}_{\omega,k} s^{(\mathrm{EUC})}_{\omega,t} + \lambda^{(\mathrm{EUC})} \sum_\omega \overline{i_{\omega,t}} f^{(\mathrm{EUC})}_{\omega,k} \left(\sum_{k'} f^{(\mathrm{EUC})}_{\omega,k'} g_{k',t} \right)^{\beta_{\mathrm{reg}}}} & (r_t \geq \tau) \\ g_{k,t} \cdot \dfrac{\sum_\omega i_{\omega,t} y_{\omega,t} f^{(\mathrm{KL})}_{\omega,k} s^{(\mathrm{KL})-1}_{\omega,t}}{\sum_\omega i_{\omega,t} f^{(\mathrm{KL})}_{\omega,k} + \lambda^{(\mathrm{KL})} \sum_\omega \overline{i_{\omega,t}} f^{(\mathrm{KL})}_{\omega,k} \left(\sum_{k'} f^{(\mathrm{KL})}_{\omega,k'} g_{k',t} \right)^{\beta_{\mathrm{reg}}}} & (r_t < \tau) \end{cases}, \tag{72}$$

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

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

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

$$N_{\omega,t} = \begin{cases} 1 & (r_t \geq \tau) \\ s_{\omega,t}^{(\mathrm{KL})-1} & (r_t < \tau) \end{cases}, \tag{75}$$

$$D_{\omega,t} = \begin{cases} s_{\omega,t}^{(\mathrm{EUC})} & (r_t \geq \tau) \\ 1 & (r_t < \tau) \end{cases}, \tag{76}$$

$$P_{\omega,l} = \begin{cases} \mu^{(\mathrm{EUC})} \sum_k f_{\omega,k}^{(\mathrm{EUC})} \sum_{\omega'} f_{\omega',k}^{(\mathrm{EUC})} h_{\omega',l} & (r_t \geq \tau) \\ \mu^{(\mathrm{KL})} \sum_k f_{\omega,k}^{(\mathrm{KL})} \sum_{\omega'} f_{\omega',k}^{(\mathrm{KL})} h_{\omega',l} & (r_t < \tau) \end{cases}. \tag{77}$$

In total, the update rules of proposed SNMF with frame-wise divergence diversity are defined as (72)–(74).

4.3 Evaluation experiment

4.3.1 Experimental conditions

To confirm the effectiveness of the proposed hybrid method with divergence diversity, I compared five methods, namely, PSNMF [22] based on KL-divergence, PSNMF based on EUC-distance, the hybrid method using SNMF with spectrogram restoration based on only EUC-distance, the hybrid method including SNMF with spectrogram restoration based on only KL-divergence, and the proposed hybrid method that switches the divergence to the optimal one framewisely. In this experiment, I used stereo signals containing four melody parts (depicted in Fig. 27) with three compositions (C1–C3) of instruments shown in Table 1. Similarly to Sect. 3.4.1, these signals were artificially generated by a MIDI synthesizer, and the observed signals $\boldsymbol{Y}$ were produced by mixing four sources with the same power. Also, these stereo signals were mixed down to a monaural format only when we evaluate the separation accuracy of PSNMF. The sources were mixed as Fig. 18, where the target source was always located in the center direction with another interfering source.

I prepared four spatially different dataset patterns of the observed signals, SP1–SP4, as shown in Table 2. In the hybrid method, many chasms were produced by directional clustering in the measures where $\theta = 45^\circ$ compared with

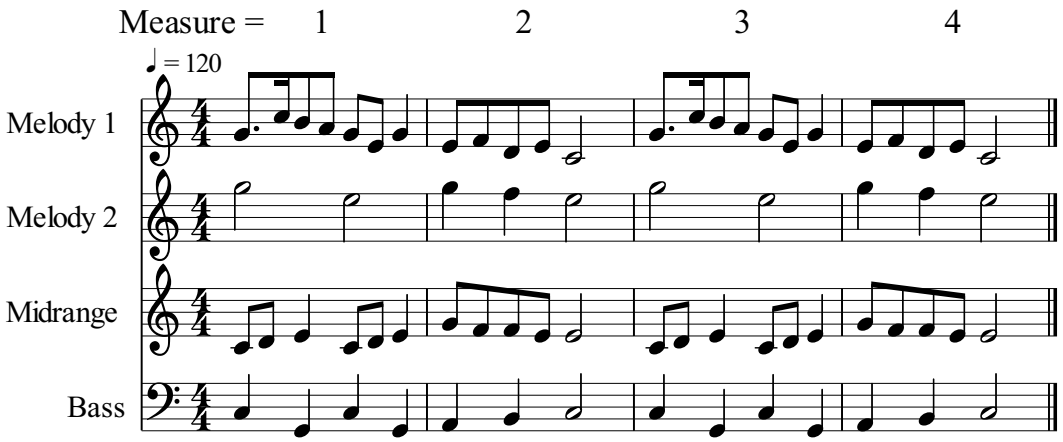

Figure 27. Scores of each part. The observed signal consists of four measures.

Table 2. Spatial conditions of each dataset

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

those of $\theta = 0°$. Therefore, we can expect that EUC-distance-based hybrid method is suitable for SP4 rather than the dataset of SP1.

In addition, I used the same MIDI sounds of the target instruments as supervision for a priori training. The threshold value for the divergence diversity, τ, was set to 20%. The other experimental conditions were the same as those in Sect. 3.4.1.

4.3.2 Experimental results

Figure 28 shows the average SDR, SIR, and SAR scores for each method and each dataset pattern, where four instruments are shuffled with 12 combinations in each of compositions C1–C3. Therefore, these results are the averages of 36 input signals. In addition, the SDR scores of PSNMF are the same for any datasets

because the input signals for PSNMF are mixed down to a monaural format. From this result, KL-divergence-based hybrid method achieves high separation accuracy for the dataset of spatial patterns SP1, SP2, and SP3 because these signals do not have much spectral chasms. On the other hand, EUC-divergence-based hybrid method achieves high separation accuracy for SP4. This dataset has many spectral chasms because the signals are always mixed with a wide panning ($\theta = 45^\circ$), which yields many chasms, and the extrapolation ability is highly required. In addition, the proposed hybrid method with frame-wise divergence diversity can always achieve better separation for any datasets regardless of the condition whether many chasms exist or not. This is because the proposed method provides the appropriate diversity of the divergence and can automatically apply the optimal divergence to each time frame.

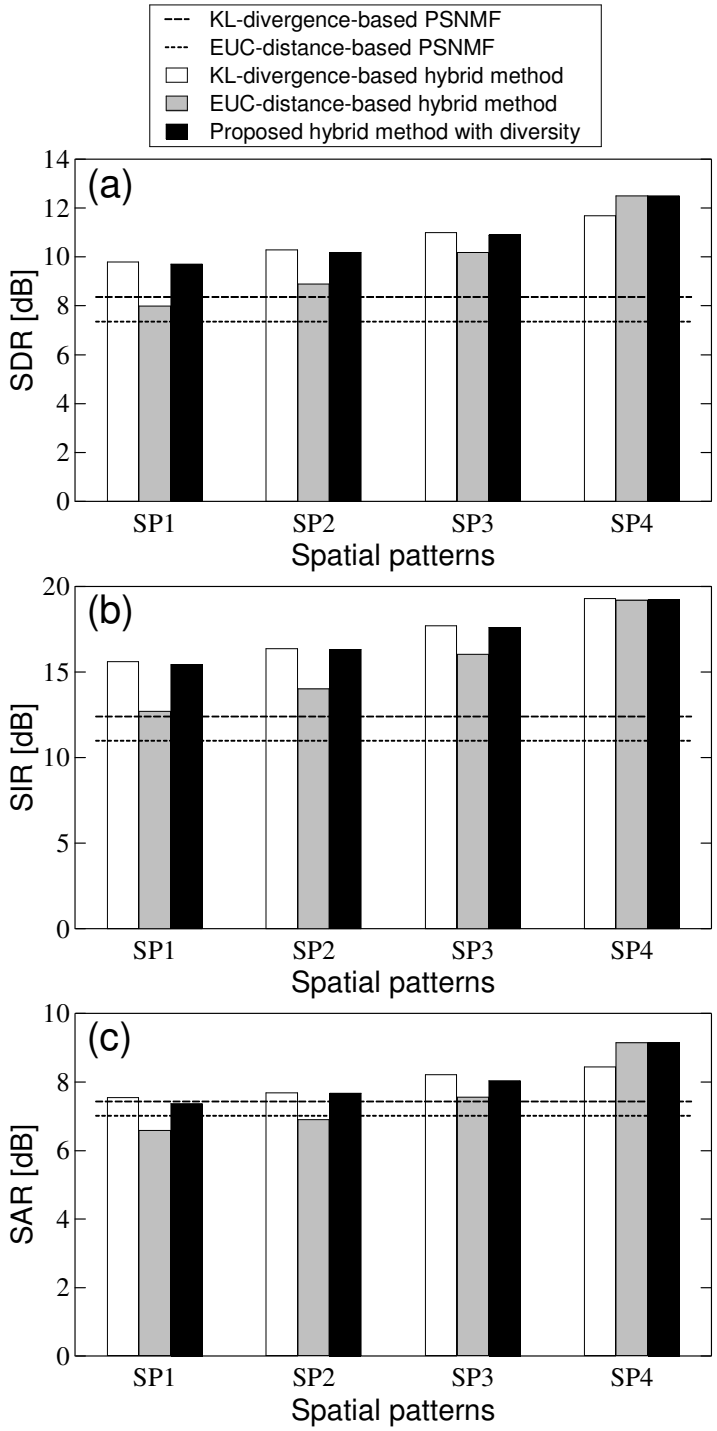

Figure 28. Average scores of each method and each spatial condition.: (a) shows SDR, (b) shows SIR, and (c) shows SAR for conventional and proposed methods.

4.4 Conclusion

In this section, first, I proposed a new divergence selection method as a improvement scheme of SNMF with spectrogram restoration and its hybrid method to separate the target signal using optimal divergence. The proposed method switches the optimal divergence in each time frame using a threshold value for the rate of the chasms to separate and extrapolate the target signal with high accuracy.

Second, I derived the update rules of proposed SNMF with divergence diversity. These update rules can switch the divergences framewisely and optimize the variable matrices, simultaneously.

Finally, I conducted the evaluation experiment to confirm the efficacy of the divergence diversity. Experimental results show that the proposed hybrid method can always achieve high separation performance under any spatial conditions.

5. Conclusion

5.1 Summary of thesis

In this thesis, I proposed a new multichannel signal separation method, i.e., a hybrid method that concatenates SNMF with spectrogram restoration after directional clustering, which reconstructs the target components lost by preceding binary masking. From theoretical 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 criterion rather than KL-divergence because of the trade-off between separation and extrapolation abilities. Based on this finding, I also proposed an improved hybrid method that switches the divergence to the optimal one framewisely. According to the results of the evaluation experiments with artificial and real-recorded signals, the proposed method is advantageous to the conventional methods in terms of robustness and separation performance.

In Sect. 3, I proposed a new SNMF with spectrogram restoration and its hybrid method for multichannel signal separation. By utilizing the index information generated from binary masking, the proposed SNMF regards the spectral chasms as unseen observations and finally reconstructs the target signal components via spectrum extrapolation using supervised bases. In other words, this SNMF can be categorized as a inpainting-based method because the deteriorated spectrogram resulting from the preceding binary masking can be recovered by the supervised basis extrapolation. In addition, a regularization term is added in the cost function to avoid extrapolation error. The theoretical analysis of the basis extrapolation ability revealed the mechanism of the marked shift of optimal divergence in SNMF with spectrogram restoration and the trade-off between separation and extrapolation abilities owing to the difference of sparseness in each divergence. Furthermore, the effectiveness of the proposed hybrid method was confirmed by the evaluation experiments with artificial and real-recorded signals.

In Section 4, I proposed an improved hybrid method. This method switches the divergence in each frame of the spectrogram according to the rate of chasms in each frame and a threshold value. Therefore, this method can be considered

as a frame-wise diversity of the divergence to achieve both optimal separation and extrapolation. Experimental results show that the proposed hybrid method with divergence diversity can always achieve high separation performance under all spatial conditions.

5.2 Future work

The following points still remain to be investigated or clarified.

- I have not analyzed other types of divergence, such as the divergence between KL-divergence and EUC-distance. I speculate that the optimal divergence strictly depends on the balance between inherent sparseness of the signal and the rate of spectral chasms generated by the preceding binary masking. Therefore, mathematical analysis of the relation between the divergence and sparseness of the signal is an important future task.

- The proposed SNMF with spectrogram restoration can be used as a postfilter for target source extraction because it reconstructs the target components lost by the preceding process using supervision. For example, we can iterate the proposed hybrid method to increase the extraction performance of the target source. The separation accuracy of the iteration method using SNMF with spectrogram restoration as a postfilter should be analyzed.

Acknowledgements

This thesis is a summary of two years of study carried out at Graduate School of Information Science, Nara Institute of Science and Technology, Japan.

I would like to express my sincere thanks to Emeritus Professor Kiyohiro Shikano and Professor Satoshi Nakamura of Nara Institute of Science and Technology, my thesis advisers, for their valuable guidance and constant encouragement.

I would also like to express my appreciation to Professor Kazushi Ikeda of Nara Institute of Science and Technology, a member of the thesis committee, for his valuable comments on the thesis.

I would especially like to express my deep gratitude to Associate Professor Hiroshi Saruwatari of Nara Institute of Science and Technology for his continuous teaching and essential advice on both technical and non technical issues. This work could not have been accomplished without his well-directed advice, helpful suggestions, and fruitful discussions with him. I have learned many valuable aspects of being a researcher from his attitude toward study and have always enjoyed conducting research with him.

I would like to express my gratitude to Visiting Associate Professor Hirokazu Kameoka of The University of Tokyo and Associate Professor Nobutaka Ono of National Institute of Informatics, for their valuable suggestions, as well as Associate Professor Tomoki Toda and Assistant Professors Hiromichi Kawanami, Sakriani Sakti, and Graham Neubig of Nara Institute of Science and Technology, and Assistant Professor Sunao Hara of Okayama University for their instructive comments.

This work could not have been achieved without the collaboration of many researchers. I especially thank Dr. Yu Takahashi, and Dr. Kazunobu Kondo, researchers at Yamaha Corporation, for their beneficial and valuable comments.

Many staff and member s of my research group have supported me in carrying out experiments and writing this thesis at Nara Institute of Science and Technology; I would especially like to express my appreciation of the valuable discussions with them on technical issues of speech signal processing and digital signal processing, and their provision of a comfortable computing environment in our laboratory. I also wish to express my deep gratitude to Ms. Toshie Nobori

and Ms. Manami Matsuda, secretaries at our laboratory, for their kind help and support in all aspects of my research.

I appreciate the opportunity to have studied with all the students in our research group at Nara Institute of Science and Technology. I thank Mr. Ryoichi Miyazaki and Ms. Fine Dwinita April, who are Ph.D. candidates at Nara Institute of Science and Technology, for useful discussions on this work.

Finally, I wish to thank all members of my family for their support over many years.

References

[1] N. Kamado, H. Nawata, H. Saruwatari, K. Shikano, "Interactive controller for audio object localization based on spatial representative vector operation," *Proc. International Workshop on Acoustic Echo and Noise Control*, 2010.

[2] H. Nawata, N. Kamado, H. Saruwatari, K. Shikano, "Automatic musical thumbnailing based on audio object localization and its evaluation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.41-44, 2011.

[3] A. J. Berkhout, "A holographic approach to acoustic control," *Journal of Audio Engineering Society*, vol.36, no.12, pp.977–995, 1988.

[4] H. Kirchhoff, S. Dixon, A. Klapuri, "Missing template estimation for user-assisted music transcription," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.26–30, 2013.

[5] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, A. Klapuri, "Automatic music transcription: Breaking the glass ceiling," *Proc. 13th International Conference on Music Information Retrieval*, 2012.

[6] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," *Proc. 16th European Signal Processing Conference*, 2008.

[7] A. Mesaros, T. Virtanen, A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," *Proc. 8th International Conference on Music Information Retrieval*, pp.375–378, 2007.

[8] M. Every, J. Szymanski, "A spectral-filtering approach to music signal separation," *Proc. 7th International Conference on Digital Audio Effects*, pp.197–200, 2004.

[9] L. Atlas, C. Janssen, "Coherent modulation spectral filtering for single-channel music source separation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.461–464, 2005.

[10] Z. Duan, Y. Zhang, C. Zhang, Z. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Trans. on Audio, Speech, and Language Processing*, vol.16, no.4, pp.766–778, 2008.

[11] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, vol.36, no.3, pp.287–314, 1994.

[12] H. Sawada, R. Mukai, S. Araki, S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. on Speech and Audio Processing*, vol.12, no.5, pp.530–538, 2004.

[13] M. Joho, H. Mathis, R. H. Lambert, "Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture," *Proc. 2nd International Conference on Independent Component Analysis and Blind Signal Separation*, pp.81–86, 2000.

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

[15] 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.

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

[17] A. Ozerov, C. Fevotte, M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.121–124, 2009.

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

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

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

[21] 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. AES 45th Conference on Applications of Time-Frequency Processing in Audio*, 2012.

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

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

[24] 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.

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

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

[27] S. Miyabe, K. Masatoki, H. Saruwatari, K. Shikano, T. Nomura, "Temporal quantization of spatial information using directional clustering for multichannel audio coding," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.261–264, 2009.

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

[29] Y. Iwao, H. Saruwatari, K. Shikano, K. Kondo, Y. Takahashi, "Stereo music signal separation combining directional clustering and nonnegative matrix factorization," *Proc. IEEE International Symposium on Signal Processing and Information Technology*, 2012.

[30] P. Smaragdis, B. Raj, "Example-driven bandwidth expansion," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.135–138, 2007.

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

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

[33] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, S. Sagayama, "Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence," *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, 2010.

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

[35] 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.

List of Publications

Jornal Papars

1. 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 Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, (in printing)

Peer Reviewed International Conference Proceedings

1. D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, "Superresolution-based stereo signal separation with regularization of supervised basis extrapolation," *Proc. 5th International Conference on 3D Systems and Applications 2013*, June S10-4, 2013.

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

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

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

5. D. Kitamura, H. Saruwatari, S. Nakamura, Y. Takahashi, K. Kondo, and H. Kameoka, "Online divergence switching for superresolution-based nonnegative matrix factorization," *Proc. 2014 RISP International Workshop*

on Nonlinear Circuits, Communications and Signal Processing, pp.485–488, March 2014.

6. T. Miyauchi, D. Kitamura, H. Saruwatari, S. Nakamura, "Depth estimation of sound images using directional clustering and activation-shared nonnegative matrix factorization," *Proc. 2014 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing*, pp.437–440, March 2014.

7. Y. murota, D. Kitamura, S. Nakai, H. Saruwatari, S. Nakamura, Y. Takahashi, and K. Kondo, "Music signal separation based on bayesian spectral amplitude estimator with automatic target prior adaptation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014 (in printing).

Technical Reports

1. D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, "Signal separation for real instruments based on supervised NMF with basis deformation," *IEICE Technical Report*, EA2012-121, vol.112, no.338, pp.13–18, March 2013 (in Japanese).

2. D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, "Importance of regularization in superresolution-based multichannel signal separation with nonnegative matrix factorization," *99th IPSJ Special Interest Group on Music and Computer*, vol.2013-MUS-99, no.14, pp.1–6, May 2013.

3. D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, "Study on optimal divergence for superresolution-based supervised nonnegative matrix factorization," *IEICE Technical Report*, EA2013-14, vol.113, no.27, pp.79–84, May 2013.

4. D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, "Music signal separation using supervised nonnegative matrix factorization

with orthogonality and maximum-divergence penalties," *28th Signal Processing Symposium*, C1-4, pp.539–544, November 2013.

5. Y. Murota, D. Kitamura, S. Nakai, H. Saruwatari, S. Nakamura, Y. Takahashi, and K. Kondo, "Postfilter-based nonnegative matrix factorization with statistical model parameter estimation," *IEICE Technical Report*, EA2013-116, vol.113, no.413, pp.75–80, January 2014.

6. T. Miyauchi, D. Kitamura, H. Saruwatari, and S. Nakamura, "Depth estimation of sound images in mixed source using activation-shared nonnegative matrix factorization," *IEICE Technical Report*, EA2013-117, vol.113, no.413, pp.81–86, January 2014 (in Japanese).

Domestic Meetings

1. D. Kitamura, H. Saruwatari, K. Shikano, K. Kondo, and Y. Takahashi, "Evaluation of separation accuracy for various real instruments based on supervised NMF with basis deformation," *2013 Spring Meeting of Acoustical Society of Japan*, 3-1-11, pp.1057-1060, March 2013.

2. D. Kitamura, H. Saruwatari, S. Nakamura, K. Kondo, and Y. Takahashi, "Divergence optimization based on trade-off between separation and extrapolation abilities in superresolution-based nonnegative matrix factorization," *2013 Autumn Meeting of Acoustical Society of Japan*, 1-1-6, pp.583-586, September 2013.

3. D. Kitamura, H. Saruwatari, S. Nakamura, K. Kondo, Y. Takahashi, and H. Kameoka, "Optimal divergence diversity for superresolution-based nonnegative matrix factorization," *2014 Spring Meeting of Acoustical Society of Japan*, 3-2-9, March 2014.

4. T. Miyauchi, D. Kitamura, H. Saruwatari, and S. Nakamura, "Automatic depth estimation of sound images using directional clutering and nonnegative matrix factorization," *2013 Autumn Meeting of Acoustical Society of Japan*, 2-1-19, pp.673–676, September 2013 (in Japanese).

5. T. Miyauchi, D. Kitamura, H. Saruwatari, and S. Nakamura, "Depth estimation of sound images in mixed source using activation-shared nonnegative matrix factorization," *2014 Spring Meeting of Acoustical Society of Japan*, 1-1-5, March 2014 (in Japanese).

6. Y. Murota, D. Kitamura, S. Nakai, H. Saruwatari, S. Nakamura, Y. Takahashi, and K. Kondo, "Experimental evaluation of postfilter-based nonnegative matrix factorization with statistical model parameter estimation," *2014 Spring Meeting of Acoustical Society of Japan*, 2-1-5, March 2014.

Awards

1. Student Award of Acoustical Society of Japan, D. Kitamura, September 2013.

2. SIP Young Researcher's Award of IEICE Technical Group on Signal Processing, D. Kitamura, February 2014.

3. Student Paper Award of 2014 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, D. Kitamura, H. Saruwatari, S. Nakamura, K. Kondo, Y. Takahashi, and H. Kameoka, March 2014.