Stereophonic Music Separation Based on Non-negative Tensor Factorization with Cepstrum Regularization

Shogo Seki[1], Tomoki Toda[2], Kazuya Takeda[1]
[1]Graduate School of Information Science, Nagoya University
[2]Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464–8601, Japan
seki.shogo@g.sp.m.is.nagoya-u.ac.jp, tomoki@itcs.nagoya-u.ac.jp, kazuya.takeda@nagoya-u.ac.jp

Abstract—**This paper presents a novel approach to stereophonic music separation based on Non-negative Tensor Factorization (NTF). Stereophonic music is roughly divided into two types; recorded music or synthesized music, which we focus on synthesized one in this paper. Synthesized music signals are often generated as linear combinations of many individual source signals with their mixing gains (i.e., time-invariant amplitude scaling) to each channel signal. Therefore, the synthesized stereophonic music separation is the underdetermined source separation problem where phase components are not helpful for the separation. NTF is one of the effective techniques to handle this problem, decomposing amplitude spectrograms of the stereo channel music signal into basis vectors and activations of individual music source signals and their corresponding mixing gains. However, it is essentially difficult to obtain sufficient separation performance in this separation problem as available acoustic cues for separation are limited. To address this issue, we propose a cepstrum regularization method for NTF-based stereo channel separation. The proposed method makes the separated music source signals follow the corresponding Gaussian mixture models of individual music source signals, which are trained in advance using their available samples. An experimental evaluation using real music signals is conducted to investigate the effectiveness of the proposed method in both supervised and unsupervised separation frameworks. The experimental results demonstrate that the proposed method yields significant improvements in separation performance in both frameworks.**

I. Introduction

Music signals are widely available through various types of music media, e.g., CDs and download services via the internet, which we can listen to using several devices, such as portable audio players, computers, and smartphones. The music signals are usually composed of many source signals, such as various instrumental sounds and vocals, and are often presented to the listener as two-channel, stereophonic signals targeting the left and right ears of the listener. An effective source separation technique for breaking up stereophonic music signals into the individual source signals is expected to be effectively used in various applications, such as music transcription [1] and extraction of vocals from music signals [2], [3].

A framework to separate mixed observation signals into individual source signals using only the mixed observation signals is known as Blind Source Separation (BSS). BSS is classified into some problems, depending on the relationship between the number of the observed signals and the number of the source signals. If the number of the source signals is larger than that of the observation signals (i.e., the number of observation channels), Independent Component Analysis (ICA) [4] is often applied to BSS. ICA is applicable to build a time-invariant linear separation filter by assuming independence among the source signals. On the other hand, it is essentially difficult to apply ICA to BSS for the stereo channel music signals because a time-invariant linear separation filter is basically ineffective if the number of the source signals is greater than the number of the observation channels. One of the effective source separation techniques in such an underdetermined condition is Non-negative Matrix Factorization (NMF) [5]–[7]. Assuming that time-frequency representations of the individual source signals are sparse, an amplitude or power spectrogram of the observation signal is modeled as a low-rank structure, i.e., a linear combination of fixed spectral patterns and corresponding time-varying weight patterns called activations. An estimate of the spectrogram of each source signal is reconstructed by using the spectral patterns and their activations corresponding to the source signal, and then, it is used to design the time-variant separation filter, such as Wiener filter.

To apply the NMF-based separation technique to the stereo music signals, it is necessary to extend a standard NMF framework for handling a single observation signal so as to also handle a stereo observation signal. This extension has been well studied as a sound source separation technique using microphone array. For instance, Multichannel NMF (MNMF) was proposed [8], [9] to make it possible to effectively use the spatial information by modeling inter-channel phase information as well as the amplitude or power spectrograms of the individual sound source signals. Furthermore, as an approach to the BSS in an overdetermined condition, Independent Low-Rank Matrix Analysis (ILRMA) [10] was proposed to explicitly use a physical constraint on the spatial mixing process. These conventional studies have shown that available acoustic cues, such as the inter-channel phase information or the special structure of a spatial mixing matrix, are effectively

ISBN 978-0-9928626-7-1 © EURASIP 2017

used to achieve good separation performance.

In this paper, we address a music source separation problem focusing on synthesized stereophonic music signals as widely used music media, which are generated as linear combinations of many individual source signals with their mixing gains (i.e., time-invariant amplitude scaling) to each channel signal. To apply the NMF-based separation technique to this underdetermined separation problem, NMF is straightforwardly extended to Nonnegative Tensor Factorization (NTF) to make it possible to model the mixing gains as well as the amplitude or power spectrograms of the individual source signals by using a tensor representation (i.e., a multidimensional array). However, it is actually difficult to achieve sufficient separation performance in this framework because available acoustic cues for separation are limited. In the synthesized stereophonic music signals, inter-frame phase information is not helpful for the source separation. Therefore, it is necessary to develop a framework making it possible to effectively use prior information as an additional cue for separation.

In order to improve separation performance of the NTF-based separation technique in the synthesized stereophonic music signals, we propose a cepstrum regularization method for the NTF-based stereo channel separation, inspired by a source separation technique using a similar idea [11]. As a prior information on the individual source signals, statistical characteristics of their timbre features are modeled as probability density functions (p.d.f.s) of their cepstral coefficients, which are trained in advance using available samples of the individual source signals. And then, the proposed method uses a novel objective function additionally consisting of negative likelihoods of the p.d.f.s for the separated source signals as a regularization term, which makes the timber features of the separated source signals close to those modeled by the p.d.f.s. The proposed method can be applied to both an unsupervised separation framework to estimate all NTF parameters and a supervised separation framework to estimate only a part of them [12]. An experimental evaluation is conducted in both supervised and unsupervised separation frameworks, demonstrating that the proposed method yields significant separation performance improvements in both frameworks.

II. MIXING PROCESS ASSUMED IN STEREO CHANNEL MUSIC SIGNAL

We assume that the observed stereo channel music signals are obtained by panning (i.e., controlling amplitude of) each composing music source signal to left and right channels and then mixing the resulting stereo channel signals of individual music source signals. In the NTF-based separation, we further assume that this mixing process is approximately applied to the amplitude/power spectral domain.

Let $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{K\times N\times C}$, $\boldsymbol{G} \in \mathbb{R}^{M\times C}$, and $\hat{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{K\times N\times M}$ represent stereo channel observation signals, a panning gain matrix, and composing music source signals in the amplitude/power spectral domain, which K, N, M, and C denote the total numbers of frequency bins, time frames, sources, and channels. Furthermore, $\hat{\boldsymbol{\mathcal{X}}}$ is decomposed to a set of basis vectors $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{K\times B\times M}$, and their corresponding activations $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{B\times N\times M}$ by using NMF, where B denotes the number of basis vectors for each composing music source signal. Each estimate of the stereo channel observation signals $\hat{s}_{knc}$ and that of the low-rank representation of the composing music source signals $\hat{x}_{knm}$ are respectively modeled as follows:

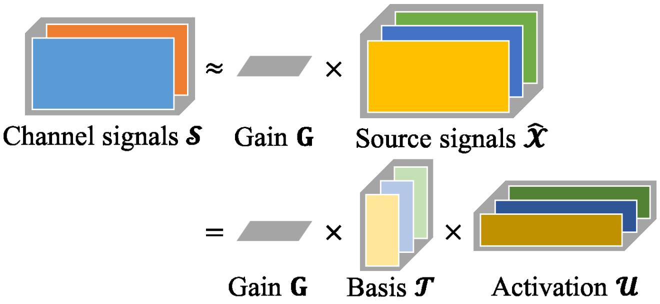
Channel signals $\boldsymbol{\mathcal{S}}$
≈
Gain $\mathbf{G}$
×
Source signals $\hat{\boldsymbol{\mathcal{X}}}$
=
Gain $\mathbf{G}$
×
Basis $\boldsymbol{\mathcal{T}}$
×
Activation $\boldsymbol{\mathcal{U}}$

Fig. 1. Frequency-independent gain NTF

$$\hat{s}_{knc} = \sum_{m} g_{mc}\hat{x}_{knm}, \tag{1}$$

$$\hat{x}_{knm} = \sum_{b} t_{kbm}u_{bnm}, \tag{2}$$

where $k \in \{1,\ldots,K\}$, $b \in \{1,\ldots,B\}$, $n \in \{1,\ldots,N\}$, $m \in \{1,\ldots,M\}$, and $c \in \{1,\ldots,C\}$ are indices of frequency bins, basis vectors, frames, sources, and channels, respectively. The variables, g_{mc}, t_{kbm}, and u_{bnm}, represent components of the parameter sets to be estimated while all of them are nonnegative. This mixing process is represented as NTF to decompose tensor-form observations into tensor-form factors. In this paper, this decomposition method is called frequency-independent gain NTF (as shown in Fig. 1) because the panning gain is fixed and independent over frequency.

III. PROPOSED METHOD: NTF WITH CEPSTRUM REGULARIZATION

The use of prior information on the composing music source signals is expected to be helpful for finding reasonable solutions of the parameter estimation in the frequency-independent gain NTF. In this paper, we propose a cepstrum regularization method for the frequency-independent gain NTF. Spectral envelope of each composing music source signal is parameterized into cepstrum, and then, its probability density is modeled with a Gaussian Mixture Model (GMM), which is trained in advance using available training data. In parameter estimation of the NTF, negative likelihood of the GMM for the separated composing music source signal is used as the regularization term. An overview of the proposed NTF with the cepstrum regularization is shown in Fig. 2.

A. Introduction of Cepstrum Regularization

The objective function with the cepstrum regularization to be minimized is defined as follows:

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{D}_{\cdot}(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\mathcal{S}}}) + \lambda\mathcal{K}(\hat{\boldsymbol{\mathcal{X}}}), \tag{3}$$

where $\mathcal{D}_{\cdot}(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\mathcal{S}}})$ is an error function between the observations and the estimates given in Eq. (1). λ is a regularization

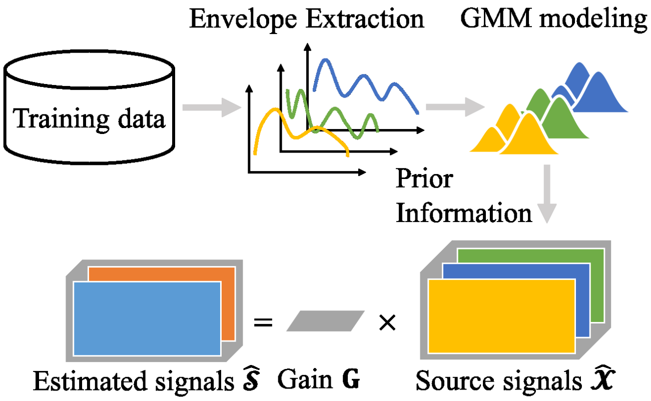
Envelope Extraction
GMM modeling
Training data
Prior
Information
=
×
Estimated signals $\hat{\boldsymbol{\mathcal{S}}}$
Gain $\mathbf{G}$
Source signals $\hat{\boldsymbol{\mathcal{X}}}$

Fig. 2. Overview of the proposed NTF with cepstrum regularization.

parameter, and $\mathcal{K}(\hat{\boldsymbol{\mathcal{X}}})$ is a cepstrum regularization term for estimates of the individual composing music source signals $\hat{\boldsymbol{\mathcal{X}}}$. The cepstrum regularization term is defined as a sum of the negative log-scaled likelihoods of the source-dependent GMM for cepstral sequences of $\hat{\boldsymbol{\mathcal{X}}}$ over all composing music source signals, which is given by

$$\mathcal{K}(\hat{\boldsymbol{\mathcal{X}}}) = \sum_{n,m} \left[-\log \sum_{p} w_{pm} \prod_{q} \mathcal{N}(E_{qnm}; \mu_{qpm}, \sigma^2_{qpm}) \right], \tag{4}$$

where E_{qnm} is the cepstrum feature of the individual estimated sources. In this paper, we use Mel-Frequency Cepstral Coefficients (MFCCs) as the cepstrum feature, which is given by

$$E_{qnm} = \sum_{r} c_{qr} \log \left| \sum_{k} f_{rk} \hat{x}_{km} \right|, \tag{5}$$

where $\boldsymbol{f} = \{f_{rk}\} \in \mathbb{R}^{R \times K}$, and $\boldsymbol{c} = \{c_{rk}\} \in \mathbb{R}^{Q \times R}$ are a R-dimensional filter-bank matrix, and an inverse cosine transform matrix, respectively. A p.d.f. of the 1st-through-Qth cepstral coefficients ($q \in \{1, \cdots, Q\}$) of each composing music source signal is modeled with the corresponding source-dependent GMM, of which parameters $\{\boldsymbol{\mu}_{pm}, \boldsymbol{\Sigma}_{pm}, w_{pm}\}$ consisting of mixture-dependent mean vectors $\boldsymbol{\mu}_{pm} = (\mu_{p1m}, \ldots, \mu_{pQm})^{\mathsf{T}}$, covariance matrices $\boldsymbol{\Sigma}_{pm} = \mathrm{diag}(\sigma^2_{p1m}, \ldots, \sigma^2_{pQm})$, and mixture component weights w_{pm}, are estimated in advance using available samples of the composing music source signal as training data. The regularization term works so that spectral envelopes of the estimated composing music source signals are similar to the desired ones, which are modeled with the source-dependent GMMs.

B. Auxiliary function design

The auxiliary function approach [13] is applied to parameter estimation. Firstly, we design an upper bound function of the error function $\mathcal{D}_{\cdot}(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\mathcal{S}}})$. In this paper, we use the KL-divergence;

$$\mathcal{D}_{\mathrm{KL}}(y|x) = y \log \frac{y}{x} - (y - x), \tag{6}$$

as the error function. From the Jensen's inequality, we obtain the following upper bound function of the error function:

$$\begin{aligned} &\mathcal{D}_{\mathrm{KL}}(\boldsymbol{\mathcal{S}}|\hat{\boldsymbol{\mathcal{S}}}) \\ &= \sum_{k,n,c} \left[s_{knc} \log \frac{s_{knc}}{\hat{s}_{knc}} - (s_{knc} - \hat{s}_{knc}) \right] \\ &\overset{c}{\leq} \sum_{k,b,n,m,c} \left[g_{mc} t_{kbm} u_{bnm} - s_{knc} \alpha_{kbnmc} \log \frac{g_{mc} t_{kbm} u_{bnm}}{\alpha_{kbnmc}} \right] \end{aligned} \tag{7}$$

where $\overset{c}{\leq}$ denotes an inequality only for the parameters to be estimated, and $\boldsymbol{\alpha} = \{\alpha_{kbnmc}\}$ is a variable satisfying $\sum_{b,m} \alpha_{kbnmc} = 1$, in which equality holds when

$$\alpha_{kbnmc} = \frac{g_{mc} t_{kbm} u_{bnm}}{\hat{s}_{knc}}. \tag{8}$$

Next, we design an upper bound function of the regularization term $\mathcal{K}^{+}(\hat{\boldsymbol{\mathcal{X}}})$. By applying similar manner as described in [14], we obtain the following upper bound function of the regularization term:

$$\begin{aligned} &\mathcal{K}(\hat{\boldsymbol{\mathcal{X}}}) \\ &\overset{c}{\leq} \sum_{r,n,m} \Bigg[A_{rnm} \left\{ \sum_{k,b} \frac{\phi^2_{rkbnm}}{f_{rk} t_{kbm} u_{bnm}} + p(\xi_{knm}) \varsigma_{rnm} + q(\xi_{rnm}) \right\} \\ &\quad - \delta_{B_{rnm}<0} |B_{rnm}| \sum_{k,b} \psi_{rkbnm} \frac{f_{rk} t_{kbm} u_{bnm}}{\psi_{rkbnm}} \\ &\quad + \delta_{B_{rnm} \geq 0} |B_{knm}| \left\{ \frac{\varsigma_{rnm}}{\zeta_{rnm}} + \log \zeta_{rnm} - 1 \right\} \Bigg], \end{aligned} \tag{9}$$

where $\boldsymbol{A} = \{A_{rnm}\}$, $\boldsymbol{B} = \{B_{rnm}\}$, and $\boldsymbol{\varsigma} = \{\varsigma_{rnm}\}$ are respectively as follows:

$$A_{rnm} = \sum_{p,q} \frac{\beta_{pnm} c^2_{qr}}{2\sigma^2_{pqm} \omega_{pqrnm}}, \tag{10}$$

$$B_{rnm} = -\sum_{p,q} \frac{\beta_{pnm} c_{qr} \gamma_{pqrnm}}{\sigma^2_{pqm} \omega_{pqrnm}}, \tag{11}$$

$$\varsigma_{rnm} = \sum_{k,b} f_{rk} t_{kbm} u_{bnm}. \tag{12}$$

Moreover, $p(\xi_{rnm})$ and $q(\xi_{rnm})$ are given by

$$p(\xi_{rnm}) = \frac{2 \log \xi_{rnm}}{\xi_{rnm}} + \frac{1}{\xi^2_{rnm}}, \tag{13}$$

$$q(\xi_{rnm}) = (\log \xi_{rnm})^2 - 2 \log \xi_{rnm} - \frac{2}{\xi_{rnm}}, \tag{14}$$

respectively, and δ_x is indicator function being 1 when the condition x is satisfied, and 0 otherwise. Equality in Eq. (9)

ISBN 978-0-9928626-7-1 © EURASIP 2017

holds when

$$\beta_{pnm} = \frac{w_{pm}\prod_q \mathcal{N}(E_{qnm};\mu_{pqm},\sigma^2_{pqm})}{\sum_{p'} w_{p'm}\prod_{q'} \mathcal{N}(E_{q'nm};\mu_{p'q'm},\sigma^2_{p'q'm})}, \quad (15)$$

$$\gamma_{pqrnm} = c_{qr}\log\varsigma_{rnm} + \omega_{pqrnm}(\mu_{pqm} - E_{qnm}), \quad (16)$$

$$\xi_{rnm} = \zeta_{rnm} = \varsigma_{rnm} = \sum_{k,b} f_{rk}t_{kbm}u_{bnm}, \quad (17)$$

$$\phi_{rkbnm} = \psi_{rkbnm} = \frac{f_{rk}t_{kbm}u_{bnm}}{\sum_{k',b'} f_{rk'}t_{k'b'm}u_{b'nm}}, \quad (18)$$

where $\boldsymbol{\beta} = \{\beta_{pnm}\}$ and $\boldsymbol{\gamma} = \{\gamma_{pqrnm}\}$ are variables satisfying $\sum_p \beta_{pnm} = 1$ and $\sum_r \gamma_{pqrnm} = \mu_{pqm}$, respectively, and $\boldsymbol{\omega} = \{\omega_{pqrnm}\}$ is an arbitrary positive constant satisfying $\sum_r \omega_{pqrnm} = 1$.

C. Parameter estimation

The auxiliary function of the objective function is represented as a sum of the upper bound functions given by Eq. (7) and Eq. (9). By partially differentiating the auxiliary function with respective parameters and setting the resulting derivatives to 0, we obtain update rules for each parameter as follows:

$$g_{mc} \leftarrow \frac{\sum_{k,b,n} s_{knc}\alpha_{kbnmc}}{t_{kbm}u_{bnm}}, \quad (19)$$

$$t_{kbm} \leftarrow \frac{-\mathsf{b}_{kbm} + \sqrt{\mathsf{b}^2_{kbm} - 4\mathsf{a}_{kbm}\mathsf{c}_{kbm}}}{2\mathsf{a}_{kbm}}, \quad (20)$$

$$u_{bnm} \leftarrow \frac{-\mathsf{e}_{bnm} + \sqrt{\mathsf{e}^2_{bnm} - 4\mathsf{d}_{bnm}\mathsf{f}_{bnm}}}{2\mathsf{d}_{bnm}}, \quad (21)$$

where a_{kbm}, b_{kbm}, c_{kbm}, d_{bnm}, e_{bnm}, and f_{bnm} are respectively given as follows:

$$\begin{aligned}\mathsf{a}_{kbm} =& \sum_{n,c} g_{mc}u_{bnm} + \lambda\sum_{r,n} A_{rnm}p(\xi_{rnm})f_{rk}u_{bnm} \\ &+ \lambda\sum_{r,n}\frac{\delta_{B_{rnm}\geq 0}|B_{rnm}|}{\zeta_{rnm}}f_{rk}u_{bnm}, \quad (22)\end{aligned}$$

$$\mathsf{b}_{kbm} = -\sum_{n,c} s_{knc}\alpha_{kbnmc} - \lambda\sum_{r,n}\delta_{B_{rnm}<0}|B_{rnm}|\psi_{rkbnm}, \quad (23)$$

$$\mathsf{c}_{kbm} = -\lambda\sum_{r,n} A_{rnm}\frac{\phi^2_{rkbnm}}{f_{rk}u_{bnm}}, \quad (24)$$

$$\begin{aligned}\mathsf{d}_{bnm} =& \sum_{k,c} g_{mc}t_{kbm} + \lambda\sum_{r,k} A_{rnm}p(\xi_{rnm})f_{rk}t_{kbm} \\ &+ \lambda\sum_{r,k}\frac{\delta_{B_{rnm}\geq 0}|B_{rnm}|}{\zeta_{rnm}}f_{rk}t_{kbm}, \quad (25)\end{aligned}$$

$$\mathsf{e}_{bnm} = -\sum_{k,c} s_{knc}\alpha_{kbnmc} - \lambda\sum_{r,k}\delta_{B_{rnm}<0}|B_{rnm}|\psi_{rkbnm}, \quad (26)$$

$$\mathsf{f}_{bnm} = -\lambda\sum_{r,k} A_{rnm}\frac{\phi^2_{rkbnm}}{f_{rk}t_{kbm}}. \quad (27)$$

IV. Experimental Evaluations

A. Experimental conditions

We conducted a music source separation experiment using real music signals. We used three songs of music data distributed from Cambridge Music Technology [15]; two of them were used for training data, and the other was used for evaluation data. Source signals of these music data were separately available, and three source signals (i.e., Bass, Drums, and Vocals) were used in the experiment. The individual source signals were separately used for the training. For the evaluation, stereophonic music signals were generated by mixing them by setting the panning gain to left and right channels to 2:1, 1:2, and 1:1 for Bass, Drums, and Vocals, respectively. All music signals were down sampled from 44.1 kHz to 16 kHz, and spectrograms were obtained with frame analysis using 32 ms window and 16 ms shift.

We evaluated the separation performance of the proposed method in both the unsupervised separation framework where all NTF parameters were estimated and the supervised separation framework where only the panning gain and activation matrices were updated while the basis vectors were set to those optimized using the training data. In evaluation, the parameters to be estimated were first updated 200 times without the cepstrum regularization, and then, they were further updated 200 times using the cepstrum regularization. The separation performance was evaluated in each setting of the regularization parameter, i.e., $\lambda = 0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2$ and 10^3, where $\lambda = 0$ was equivalent to the NTF-based separation without the cepstrum regularization updating the parameters 400 times. In order to reduce the effect of random parameter initialization on the separation performance, the separation process was conducted ten times by changing an initial setting in each condition. The number of basis vectors was set to 50 for each music source signal. A time sequence of the MFCCs was extracted from each source signal, and then, the source-dependent GMM was trained using it. MFCC orders and the number of mixture components were determined through our preliminary experiment.

As the performance measurements, Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifact Ratio (SAR) of the estimated stereo channel music source signals were calculated using BSS EVAL toolbox [16]. These measurements were calculated in each channel, and then, they were averaged over two channels.

B. Experimental Results

Figure 3 shows results of SDR, SIR, and SAR in the unsupervised and supervised separation. In each figure, the horizontal axis shows a setting of the regularization parameter and the vertical axis shows each performance measurement. Separation performance of each source signal is shown separately in the figure. Note that results of the NTF without the cepstrum regularization are shown as $\lambda = 0$.

We can see that the proposed cepstrum regularization yields significant performance improvements in both unsupervised

ISBN 978-0-9928626-7-1 © EURASIP 2017

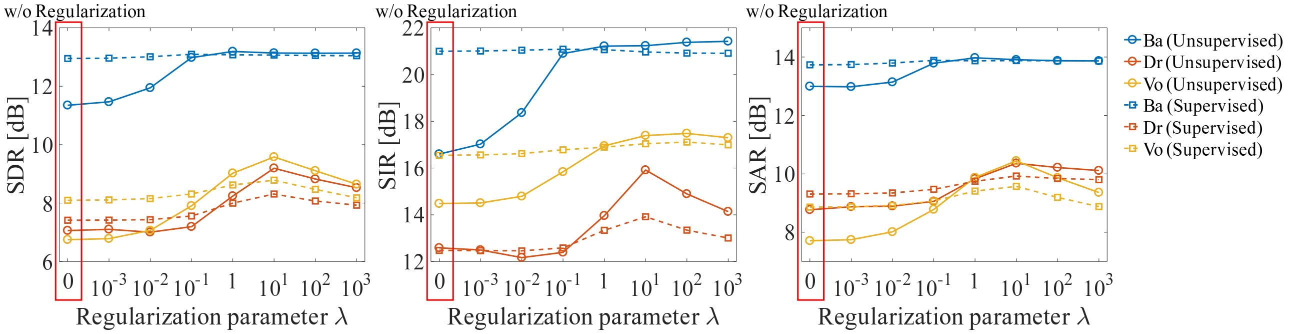

Fig. 3. Separation performance for each source signal in unsupervised and supervised separation. The cepstrum regularization is not used if the regularization parameter λ set to 0.

and supervised separation frameworks by suitably setting the regularization parameter λ to around 1 to 10^2. In such a suitable setting, we can also see that the unsupervised separation performance outperforms the supervised separation performance. On the other hand, if the cepstrum regularization is not used (i.e., $\lambda = 0$), the unsupervised separation performance significantly degrades and it becomes worse than the supervised separation performance. These results suggest that 1) the supervised separation performance is limited because the basis vectors are strongly affected by acoustic mismatches between training data and evaluation data, 2) the update of the basis vectors is helpful for compensating those mismatches but it is difficult to be achieved in the standard NTF-based separation, and 3) the proposed NTF-based separation with the cepstral regularization is capable of effectively updating the basis vectors as well and yielding significant improvements in separation performance. We can also see that the separation performance strongly depends on the individual source signals and the proposed method is more effective for the source signals causing relatively low separation performance.

V. Conclusions

In this paper, we have proposed a method for synthesized stereophonic music separation based on Nonnegative Tensor Factorization (NTF) with cepstrum regularization. The proposed method makes it possible to consider statistical characteristics of timbre features of individual source signals as prior information for separation. The cepstrum regularization works as a soft constraint to update the NTF parameters including basis vectors, and therefore, acoustic mismatches between training data and evaluation data are handled well. From the experimental results, it has been demonstrated that the proposed method yields significant improvements in separation performance compared to the standard NTF-based method without the cepstrum regularization. Furthermore, it has also been demonstrated that the proposed method is effective for both an unsupervised separation framework and a supervised separation framework.

We plan to further investigate the effectiveness of the proposed method for various music sources and also improve its separation performance by enhancing the individual source-dependent models so as to accurately model acoustic characteristics of the individual music source signals. In addition, we also plan to study an optimization method of the regularization parameter that strongly affects separation performance.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grant Number 17H01763.

References

[1] Paris Smaragdis and Judith C Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. of WASPAA*, 2003, pp. 177–180.

[2] Shankar Vembu and Stephan Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. of ISMIR*, 2005, pp. 337–344.

[3] Yukara Ikemiya, Kazuyoshi Yoshii, and Katsutoshi Itoyama, "Singing voice analysis and editing based on mutually dependent f0 estimation and source separation," in *Proc. of ICASSP*, 2015, pp. 574–578.

[4] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, *Independent component analysis*, vol. 46, John Wiley & Sons, 2004.

[5] Daniel D Lee and H Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[6] Hirokazu Kameoka, Nobutaka Ono, Kunio Kashino, and Shigeki Sagayama, "Complex nmf: A new sparse representation for acoustic signals," in *Proc. of ICASSP*, 2009, pp. 3437–3440.

[7] Paris Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. of ICA*, 2004, pp. 494–499.

[8] Alexey Ozerov and Cédric Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[9] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. on ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[10] Daichi Kitamura, Nobutaka Ono, Hiroshi Sawada, Hirokazu Kameoka, and Hiroshi Saruwatari, "Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model," in *Proc. of ICASSP*, 2015, pp. 276–280.

[11] Li Li, Hirokazu Kameoka, Takuya Higuchi, and Hiroshi Saruwatari, "Semi-supervised joint enhancement of spectral and cepstral sequences of noisy speech," in *Proc. of Interspeech*, 2016, pp. 3753–3757.

[12] Paris Smaragdis, Raj Bhiksha, and Shashanka Madhusudana, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. of ICA*, 2007, pp. 414–421.

[13] Daniel D Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Proc. of NIPS*, 2001, pp. 556–562.

[14] Hirokazu Kameoka, Masahiro Nakano, Kazuki Ochiai, Yutaka Imoto, Kunio Kashino, and Shigeki Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," in *Proc. of ICASSP*, 2012, pp. 5365–5368.

[15] "Cambridge music technology," http://cambridge-mt.com/ms-mtk.htm, Accessed: 2017-02-17.

[16] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

ISBN 978-0-9928626-7-1 © EURASIP 2017