SHIFTED AND CONVOLUTIVE SOURCE-FILTER NON-NEGATIVE MATRIX FACTORIZATION FOR MONAURAL AUDIO SOURCE SEPARATION

Tomohiko Nakamura[†] *and Hirokazu Kameoka*[†,‡]

[†]Graduate School of Information Science and Technology, The University of Tokyo.
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
[‡]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation.
3-1, Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198, Japan

ABSTRACT

This paper proposes an extension of non-negative matrix factorization (NMF), which combines the shifted NMF model with the source-filter model. Shifted NMF was proposed as a powerful approach for monaural source separation and multiple fundamental frequency (F_0) estimation, which is particularly unique in that it takes account of the constant inter-harmonic spacings of a harmonic structure in log-frequency representations and uses a shifted copy of a spectrum template to represent the spectra of different F_0s. However, for those sounds that follow the source-filter model, this assumption does not hold in reality, since the filter spectra are usually invariant under F_0 changes. A more reasonable way to represent the spectrum of a different F_0 is to use a shifted copy of a harmonic structure template as the excitation spectrum and keep the filter spectrum fixed. Thus, we can describe the spectrogram of a mixture signal as the sum of the products between the shifted copies of excitation spectrum templates and filter spectrum templates. Furthermore, the time course of filter spectra represents the dynamics of the timbre, which is important for characterizing the feature of an instrument sound. Thus, we further incorporate the non-negative matrix factor deconvolution (NMFD) model into the above model to describe the filter spectrogram. We derive a computationally efficient and convergence-guaranteed algorithm for estimating the unknown parameters of the constructed model based on the auxiliary function approach. Experimental results revealed that the proposed method outperformed shifted NMF in terms of the source separation accuracy.

Index Terms— Audio source separation, Shifted non-negative matrix factorization, Shift-invariant probabilistic latent component analysis, Source-filter theory

1. INTRODUCTION

One major approach to monaural source separation involves applying non-negative matrix factorization (NMF) to an observed magnitude (or power) spectrogram interpreted as a non-negative matrix [1]. While many variants and extensions of NMF have been applied to spectrograms with linear frequency resolution given by the short-time Fourier transform (STFT), spectrograms with log-frequency resolution obtained with the continuous wavelet transform (CWT) are particularly well suited to music data in the sense that the fundamental frequencies (F_0s) of the tones in music are geometrically spaced [2–7]. In particular, shifted NMF [2], also known as the shift-invariant probabilistic latent component analysis (PLCA) [3], has proved successful for monaural source separation and multiple F_0 estimation tasks [8–10]. The uniqueness of this method lies in that it takes account of the constant inter-harmonic spacings of a harmonic structure in log-frequency representations and uses shifted copies of a spectrum template to represent the spectra of different F_0s where the shape of each spectrum template is assumed to be associated with an individual instrument. It should be noted that this idea was first introduced in [11, 12] to develop a method for multiple F_0 estimation. However, the above assumption does not hold in real situations. Fig. 1 shows two examples of the spectra produced by the same instrument with different F_0s. As Fig. 1 shows, the relative energies of the harmonic partials of the two spectra appear to be completely different, so that it is difficult to represent them only with a single template. To address this mismatch, previous work proposed using multiple templates to represent the spectra of the same instrument [2, 8–10].

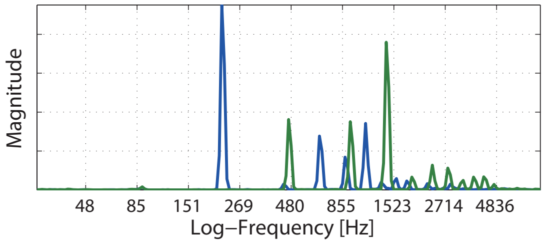

Fig. 1: Two spectra of clarinet sounds at different pitches.

To construct a more reasonable model, we focus on the fact that the generating processes of many instrument sounds can be explained fairly well by the source-filter theory. Specifically, we assume that the spectrum of an instrument sound is given by the product of the excitation and filter spectra, where the “excitation” source represents a vibrating object such as a violin string and “filter” refers to the resonance structure of the instrument body, which is usually invariant under F_0 changes. While most of the NMF variants incorporating the source-filter model such as [13–17] were designed to model STFT spectrograms, we consider modeling CWT spectrograms so as to utilize the shift-invariant property of the excitation

This study was supported by JSPS KAKENHI Grant Number 15J09992 and JSPS Grant-in-Aid for Young Scientists B Grant Number 26730100.

spectrum. We can thus describe the spectrum of each instrument at a particular time as the product of a shifted copy of an excitation spectrum template and a filter spectrum.

As regards the filter spectrum, if it can be assumed to be constant over time, we can use a single spectrum template for each instrument. However, for some instruments such as a singing voice, the time course of the filter spectrum represents the dynamics of the timbre, which is important for characterizing the feature of an instrument. We thus consider it reasonable to represent the filter spectrogram using a short-range spectrogram rather than a single-frame spectrum as the template. This can be achieved by introducing the convolutive NMF model [18] to express the filter spectrogram of each instrument.

In summary, this paper proposes modeling the CWT spectrogram of a polyphonic music as the sum of components each described by the product of the excitation and filter spectrogram, where the excitation and filter spectrograms are respectively expressed using the shifted NMF model and the convolutive NMF model. We further derive a convergence-guaranteed iterative algorithm for minimizing the difference between an observed spectrogram and the present model based on the auxiliary function approach [19, 20].

2. SHIFTED AND CONVOLUTIVE SOURCE-FILTER NON-NEGATIVE MATRIX FACTORIZATION

2.1. Spectrogram Model

Let us denote the indices of log-frequency and time by $l = 0, \dots, L-1$ and $m = 0, \dots, M-1$, respectively. A spectrogram of an audio signal that follows the source-filter model can be described as the product of an excitation spectrogram and a filter spectrogram. By using the fact that the inter-harmonic spacings of a harmonic structure are constant in the log-frequency domain, we use a shifted copy of an excitation spectrum template to represent the excitation spectrum of each F_0. The excitation spectrogram $\tilde{X}^{(\mathrm{ex})}_{l,m,k} \geq 0$ of source excitation $k(= 0, \cdots, K-1)$ is modeled as the convolution of an excitation spectrum template $S_{k,l} \geq 0$ and time-varying gains $U^{(\mathrm{ex})}_{k,p,m} \geq 0$, i.e. $\tilde{X}^{(\mathrm{ex})}_{l,m,k} = \sum_{p \in \mathcal{P}} S_{k,l-p} U^{(\mathrm{ex})}_{k,p,m}$, where p is the frequency shift index and $\mathcal{P}$ is the set of possible frequency shifts. By abuse of notation, we understand that $S_{k,l-p} = 0$ unless $0 \leq l - p \leq L - 1$.

On the other hand, we describe the filter spectrogram in a similar manner to NMFD [18] and NMF-2D [4] to capture the dynamics of the timbre, which is important for characterizing the feature of an instrument sound. The spectrogram $\tilde{X}^{(\mathrm{filt})}_{l,m,r} \geq 0$ of filter $r(= 0, \cdots, R-1)$ is represented by a time convolution of a short-range spectrogram $F_{r,l,\tau} \geq 0$ with time-varying gains $U^{(\mathrm{filt})}_{r,m} \geq 0$, i.e. $\tilde{X}^{(\mathrm{filt})}_{l,m,r} = \sum_{\tau=0}^{M^{(\mathrm{tap})}-1} F_{r,l,\tau} U^{(\mathrm{filt})}_{r,m-\tau}$, where $\tau = 0, \cdots, M^{(\mathrm{tap})} - 1$ is the time shift index and $M^{(\mathrm{tap})}$ is the tap size of the short-range spectrograms. By abuse of notation, we understand that $U^{(\mathrm{filt})}_{r,m-\tau} = 0$ unless $0 \leq m - \tau \leq M - 1$.

As we want the magnitude spectra of filters to be smooth and non-negative in the log-frequency domain, we parameterize $F_{r,l,\tau}$ by N envelope kernels $G_{l,n} \geq 0$ and their mixture weights $W_{r,n,\tau} \geq 0$ that satisfies $\sum_{n,\tau} W_{r,n,\tau} = 1$:

$$F_{r,l,\tau} = \sum_{n} W_{r,n,\tau} G_{l,n}, \tag{1}$$

$$G_{l,n} := \frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(\omega_l - \rho_n)^2}{2\nu^2}} \tag{2}$$

where $n = 0, \cdots, N-1$ is the index of envelope kernel and $\omega_l \in (0, \pi]$ is the normalized angular frequency corresponding to the lth log-frequency. The kernel $G_{l,n}$ for $n \geq 1$ is identical to a normal distribution of a normalized angular frequency with mean ρ_n and variance ν^2.

Multiple excitation and filter spectra can be used for an instrument to describe complex spectral changes, but we hereafter assign an excitation spectrum and a filter spectrum to each instrument for the simplicity. By putting $U_{k,r,p,m-\tau} = U^{(\mathrm{ex})}_{k,p,m} U^{(\mathrm{filt})}_{k,r,p,m-\tau}$ and treating $U_{k,r,p,m-\tau}$ itself as a parameter, the spectrogram of an instrument sound associated with source excitation k and filter r can be written as

$$\tilde{X}_{l,m,k,r} = \sum_{p,\tau} F_{r,l,\tau} S_{k,l-p} U_{k,r,p,m-\tau}. \tag{3}$$

Assuming the additivity of magnitude spectrograms as with conventional NMFs, the observed spectrogram can be represented as

$$X_{l,m} = \sum_{k,r} \tilde{X}_{l,m,k,r}. \tag{4}$$

To avoid the indeterminacy in scaling, we put $\sum_l S_{k,l} = 1$ for all k.

Although a model similar to the above has been mentioned in [21], the temporal dynamics was not incorporated into the source-filter model in the literature. Any experimental evaluation was not given in the literature and the incorporation of the source-filter model into shifted NMF has yet been validated. We will thus confirm the efficacy of the incorporation of the source-filter model in Sec. 4.

2.2. Formulation

For a given magnitude spectrogram $Y := \{Y_{l,m}\}_{l,m}$, we would like to find the parameters $S := \{S_{k,p}\}_{k,p}$, $W := \{W_{r,n,\tau}\}_{r,n,\tau}$ and $U := \{U_{k,r,l,m}\}_{k,r,l,m}$ of the proposed model such that minimizes

$$\mathcal{L}_*(S, W, U) = \sum_{l,m} D_*(Y_{l,m} \| X_{l,m}) + \mathcal{R}_*(U). \tag{5}$$

The first term of Eq. (5) is a goodness-of-fit measure between Y and $X := \{X_{l,m}\}_{l,m}$. How to define the measure is very important since it corresponds to an assumption to the statistical nature of observed data. If we define D_{I} as the generalized Kullback-Leibler divergence (a.k.a I divergence), it implicitly assumes that $Y_{l,m}$ follows a Poisson distribution with mean $X_{l,m}$:

$$D_{\mathrm{I}}(Y_{l,m} \| X_{l,m}) = Y_{l,m} \ln \frac{Y_{l,m}}{X_{l,m}} - Y_{l,m} + X_{l,m}. \tag{6}$$

From this fact, it is known that minimizing $\sum_{l,m} D_{\mathrm{I}}(Y_{l,m} \| X_{l,m})$ with respect to $X_{l,m}$ amounts to the maximum likelihood estimation of $X_{l,m}$. This measure is frequently used in conventional NMF algorithms and has been confirmed to work well for audio source separation empirically. Another commonly-used measure is the Itakura-Saito divergence D_{IS}:

$$D_{\mathrm{IS}}(Y^2_{l,m} \| X_{l,m}) = \frac{Y^2_{l,m}}{X_{l,m}} - \ln \frac{Y^2_{l,m}}{X_{l,m}} - 1. \tag{7}$$

This corresponds to the assumption that an observed complex spectrogram follows a circularly-symmetric complex normal distribution with mean zero and variance $X_{l,m}$, in which $X_{l,m}$ can be interpreted as a model of a power spectral density of the observed signal.

The second term $\mathcal{R}_*(U)$ is a regularizer for U. In popular and classical western music, the number of pitches occurred in a musical piece and the number of times each note is performed are usually limited, and so inducing the sparsity of U would facilitate the source separation. To reflect it, we can design the regularizer in analogy to the Bayesian modeling. The conjugate prior of the Poisson distribution is a gamma distribution $\mathrm{Gam}(x;\alpha,\beta) \propto x^{\alpha-1}e^{-\beta x}$ and thus we design the regularizer $\mathcal{R}_{\mathrm{I}}(U)$ for D_{I} as

$$\mathcal{R}_{\mathrm{I}}(U) = \sum_{k,r,p,m} \left\{-(\alpha^{(\mathrm{I})} - 1)\ln U_{k,r,p,m} + \beta^{(\mathrm{I})} U_{k,r,p,m}\right\}, \tag{8}$$

where $\alpha^{(\mathrm{I})} > 0$ and $\beta^{(\mathrm{I})} > 0$ are associated with the shape and rate parameters of a gamma distribution, respectively. Similarly, the conjugate prior of a circularly-symmetric complex normal distribution with known mean and unknown variance is an inverse gamma distribution $\mathrm{InvGam}(x;\alpha,\beta) \propto x^{-\alpha-1}e^{-\beta/x}$, and we design the regularizer for D_{IS} as

$$\mathcal{R}_{\mathrm{IS}}(U) = \sum_{k,r,p,m} \left\{(\alpha^{(\mathrm{IS})} + 1)\ln U_{k,r,p,m} + \frac{\beta^{(\mathrm{IS})}}{U_{k,r,p,m}}\right\}, \tag{9}$$

where $\alpha^{(\mathrm{IS})} > 0$ and $\beta^{(\mathrm{IS})} > 0$ are associated with the shape and scale parameters of an inverse gamma distribution, respectively. The less $\alpha^{(\mathrm{I})}$ ($\alpha^{(\mathrm{IS})}$), the more sparse U tends to become.

3. PARAMETER ESTIMATION ALGORITHMS BASED ON AUXILIARY FUNCTION APPROACH

We first derive a parameter estimation algorithm for the I divergence. Since $\mathcal{L}_{\mathrm{I}}(S,W,U)$ involves summations over k, r, p, τ and n in the logarithmic function, the current minimization problem is difficult to solve analytically. However, we can develop a computationally efficient algorithm for finding a locally optimal solution based on the auxiliary function approach [19, 20, 22]. The first step to apply the auxiliary function approach, is to define an upper bound function for the objective function $\mathcal{L}(S,W,U)$, arranged as $\mathcal{L}^{+}(S,W,U,\Lambda)$, such that $\mathcal{L}(S,W,U) = \min_{\Lambda}\mathcal{L}^{+}(S,W,U,\Lambda)$. We call Λ an auxiliary variable and $\mathcal{L}^{+}(S,W,U,\Lambda)$ an auxiliary function. If we can construct $\mathcal{L}^{+}(S,W,U,\Lambda)$, $\mathcal{L}(S,W,U)$ is non-increasing under the updates $\{S,W,U\} \leftarrow \underset{S,W,U}{\mathrm{argmin}}\,\mathcal{L}^{+}(S,W,U,\Lambda)$ and $\Lambda \leftarrow \underset{\Lambda}{\mathrm{argmin}}\,\mathcal{L}^{+}(S,W,U,\Lambda)$.

Since the logarithmic function is a concave function, we can obtain an upper bound function by invoking the Jensen's inequality:

$$\begin{aligned} -Y_{l,m}\ln X_{l,m} \leq -\,Y_{l,m}\sum_{k,r,p,\tau,n}\lambda_{l,m,k,r,p,\tau,n}\Big(&\ln S_{k,l-p} + \ln W_{r,n,\tau} \\ &+ \ln G_{l,n} + \ln U_{k,r,p,m-\tau} - \ln\lambda_{l,m,k,r,p,\tau,n}\Big) \end{aligned} \tag{10}$$

where $\lambda_{l,m,k,r,p,\tau,n} \geq 0$ is an auxiliary variable such that $\sum_{k,r,p,\tau,n}\lambda_{l,m,k,r,p,\tau,n} = 1$ for all l and m. The equality holds if and only if

$$\lambda_{l,m,k,r,p,\tau,n} = \frac{S_{k,l-p}W_{r,n,\tau}G_{l,n}U_{k,r,p,m-\tau}}{X_{l,m}}. \tag{11}$$

The auxiliary function can thus be written as

$$\begin{aligned} \mathcal{L}_{\mathrm{I}}^{+}(S,W,U,\Lambda) \underset{c}{=} -\sum_{l,m}Y_{l,m}\sum_{k,r,p,\tau,n}\lambda_{l,m,k,r,p,\tau,n}\Big(&\ln S_{k,l-p} + \ln W_{r,n,\tau} \\ &+ \ln U_{k,r,p,m-\tau} - \ln\lambda_{l,m,k,r,p,\tau,n}\Big) + \sum_{l,m}X_{l,m} \\ &- \sum_{k,r,p,m}\left\{(\alpha^{(\mathrm{I})} - 1)\ln U_{k,r,p,m} - \beta^{(\mathrm{I})}U_{k,r,p,m}\right\} \end{aligned} \tag{12}$$

where $\underset{c}{=}$ denotes the equality up to constant terms and $\Lambda := \{\lambda_{l,m,k,r,p,\tau,n}\}_{l,m,k,r,p,\tau,n}$. By setting the partial derivatives of $\mathcal{L}^{+}(S,W,U,\Lambda)$ with respect to S, W and U at zeros and substituting Eq. (11) into Λ, we can derive the following update equations:

$$S_{k,l'} \leftarrow S_{k,l'}\frac{\sum_{l,m}\frac{Y_{l,m}}{X_{l,m}}\sum_{r,\tau}F_{r,l,\tau}U_{k,r,l-l',m-\tau}}{\sum_{l,m,r,\tau}F_{r,l,\tau}U_{k,r,l-l',m-\tau}}, \tag{13}$$

$$W_{r,n,\tau} \leftarrow W_{r,n,\tau}\frac{\sum_{l,m}\frac{Y_{l,m}}{X_{l,m}}\sum_{p,k}G_{l,n}S_{k,l-p}U_{k,r,p,m-\tau}}{\sum_{l,m,p,k}G_{l,n}S_{k,l-p}U_{k,r,p,m-\tau}}, \tag{14}$$

$$U_{k,r,p,m'} \leftarrow \frac{\sum_{l,m}\frac{Y_{l,m}}{X_{l,m}}\sum_{p,k}F_{r,l,m-m'}S_{k,l-p}U_{k,r,p,m'} + \alpha^{(\mathrm{I})} - 1}{\sum_{l,m,p,k}F_{r,l,m-m'}S_{k,l-p} + \beta^{(\mathrm{I})}}, \tag{15}$$

The update rules of S and W are followed by normalization such that $\sum_{l}S_{k,l} = 1$ for all k and $\sum_{n,\tau}W_{r,n,\tau} = 1$ for all r. It is important to note that once the initial values of W and S are set to be non-negative, the multiplicative update equations ensures the non-negativity of the entries of W and S. Since the non-negativity of U does not hold, we can ensure it by simply performing $U_{k,r,p,m} \leftarrow \max\{0, U_{k,r,p,m}\}$ at each update. One may think that the update equations contain time-consuming convolutions and correlations and would require a long computation time. However, we can invoke the fast Fourier transform (FFT) to calculate the convolutions and correlations, and they are not time-consuming in practice.

Similarly to the above, we can construct an auxiliary function for the IS divergence as with [23]. We here omit details of the derivation of an auxiliary function due to limitations of space. Update equations can be derived as

$$S_{k,l'} \leftarrow S_{k,l'}\sqrt{\frac{\sum_{l,m,r,\tau}\frac{Y_{l,m}^2}{X_{l,m}^2}F_{r,l,\tau}U_{k,r,p,m}}{\sum_{l,m,r,\tau}\frac{F_{r,l,\tau}U_{k,r,p,m}}{X_{l,m}}}}, \tag{16}$$

$$W_{r,n,\tau} \leftarrow W_{r,n,\tau}\sqrt{\frac{\sum_{l,m,k,p}\frac{Y_{l,m}^2}{X_{l,m}^2}S_{k,l-p}G_{l,n}U_{k,r,p,m}}{\sum_{l,m,k,p}\frac{S_{k,l-p}G_{l,n}U_{k,r,p,m}}{X_{l,m}}}}, \tag{17}$$

$$U_{k,r,p,m'} = \frac{A_{k,r,p,m'}}{\sqrt{\left(\frac{\alpha^{(\mathrm{IS})}+1}{2}\right)^2 + B_{k,r,p,m'}A_{k,r,p,m'}} + \frac{\alpha^{(\mathrm{IS})}+1}{2}} \tag{18}$$

where

$$A_{k,r,p,m'} = \sum_{l,m}\frac{Y_{l,m}^2}{X_{l,m}^2}F_{r,l,m-m'}S_{k,l-p}U_{k,r,p,m'^2} + \beta^{(\mathrm{IS})}. \tag{19}$$

$$B_{k,r,p,m'} = \sum_{l,m}\frac{S_{k,l-p}F_{r,l,m-m'}}{X_{l,m}}. \tag{20}$$

4. EXPERIMENTS

To evaluate the proposed algorithms, we conducted a supervised source separation experiment. For the convenience, we call the proposed algorithm with the I divergence criterion (the IS divergence

Table 1: Average SDR improvements, SIR improvements and SARs with standard errors obtained with the proposed algorithms (*I-SNMFwSF* and *IS-SNMFwSF*) and the shifted NMFs (*I-SNMF* and *IS-SNMF*).

Algorithm	SDR improvements	SIR improvements	SARs
I-SNMFwSF $(0.6, 1.0 \times 10^{-10})$	$\mathbf{6.01 \pm 0.58}$	$\mathbf{11.06 \pm 1.09}$	3.91 ± 0.64
IS-SNMFwSF $(1.0, 0.6)$	4.77 ± 0.29	8.98 ± 0.40	3.46 ± 0.50
I-SNMF $(1.0, 0.4)$	3.96 ± 0.44	9.58 ± 0.69	1.79 ± 0.85
IS-SNMF $(0.4, 1.0)$	2.81 ± 0.51	6.00 ± 0.61	$\mathbf{4.09 \pm 0.94}$

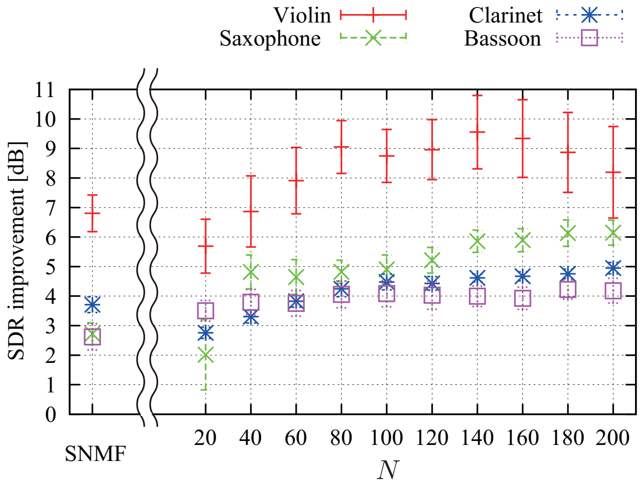

Fig. 2: Average SDR improvements and standard errors obtained with the proposed algorithm (*I-SNMFwSF*) and *I-SNMF* for each musical instrument. "SNMF" corresponds to *I-SNMF*.

criterion) *I-SNMFwSF* (*IS-SNMFwSF*, respectively). For comparison, we employed shifted NMF with the I divergence criterion (*I-SNMF*) and that with the IS divergence criterion (*IS-SNMF*). While the original shifted NMF [2] does not contain any terms inducing the sparsity of parameters, we found that $\mathcal{R}_*(U)$ improved SDRs in the experiment and we here used $\mathcal{R}_*(U)$.

The experimental data was the Bach10 dataset [24], which consists of audio recordings of ten four-part chorales by J. S. Bach. Each recording is a mixture of violin, clarinet saxophone and bassoon performances, which correspond to the soprano, alto, tenor and bass parts of each musical piece, respectively. Audio recordings of individual parts are also contained in the dataset. All recordings were monaural and downsampled to 16 kHz. Magnitude spectrograms were computed with the fast approximate CWT algorithm [25, 26]. The center frequencies ranged from 27.5 to 7902 Hz with 100/3 cent interval and the log-normal wavelet [27] was used as an analyzing wavelet. The wavelet has a Gaussian shape with a common variance in the log-frequency domain, and we set a parameter corresponding to the standard deviation of the Gaussian as a one fifth of a semitone interval.

We first trained S and W of the proposed models and basis spectra of the shifted NMFs with the audio recordings of individual parts of the five musical pieces (training data), and then performed source separation on the audio recordings of the other five musical pieces (test data). With the proposed algorithms, a pair of a source excitation and a filter was trained for each instrument, and a total of four pairs of a source and a filter were used for the separation. With the shifted NMFs, one basis spectrum was assigned to each instrument and a total of four basis spectra were used for the separation. For each test data, we designed a soft time-frequency mask as $\tilde{X}_{l,m,k,r}/X_{l,m}$ to obtain separated audio signals of the sources. The proposed methods and the shifted NMFs ran for 100 iterations both in the training and test stages. As $\alpha^{(\mathrm{I})}$ or $\alpha^{(\mathrm{IS})}$, we use $\alpha^{(\mathrm{train})} = 1.0 \times 10^{-10}, 0.2, 0.4, 0.6, 0.8, 1.0$ for the training data and $\alpha^{(\mathrm{test})} = 1.0 \times 10^{-10}, 0.2, 0.4, 0.6, 0.8, 1.0$ for the test data. The other parameters were set as follows: $\beta^{(\mathrm{I})} = \beta^{(\mathrm{IS})} = 1.0 \times 10^{-10}$, $M^{(\mathrm{tap})} = 1$, $N = 140$, $\nu = \pi/(2N-2)$ and $\rho_n = \pi n/(N-1)$ for $n = 0, \cdots, N-1$.

Table 1 shows SDR improvements, SIR improvements and SARs obtained with all algorithms for all data. SDRs, SIRs and SARs were computed with the BSSEval toolbox [28]. The pairs of two values below the algorithm names are $(\alpha^{(\mathrm{train})}, \alpha^{(\mathrm{test})})$, which provided the highest average SDR improvement for each algorithm. These results show that the incorporation of the source-filter model improves the source separation accuracy in the CWT domain.

We also conducted an experiment to examine the effect of N of the proposed algorithms. Fig. 2 displays average SDR improvements and standard errors obtained with *I-SNMFwSF* and *I-SNMF* for individual musical instruments, where each algorithm ran with the same $(\alpha^{(\mathrm{train})}, \alpha^{(\mathrm{test})})$ as in the above. *I-SNMFwSF* with $100 \leq N \leq 160$ provided significantly higher SDR improvements for all musical instruments compared to *I-SNMF*. We found that the best N was different for each musical instrument, and so exploring the best Ns for other musical instruments and classifying them may be required for practical use.

5. CONCLUSION

This paper has developed a new source separation method by incorporating the source-filter model into shifted NMF. With the proposed model, the observed spectrogram is represented by a product of excitation and filter spectrograms. The excitation spectrogram is described with shifted NMF to exploit the constant inter-harmonic spacings of a harmonic structure in the log-frequency domain, and the filter spectrogram is modeled by NMFD to represent temporal dynamics of timbre. We have derived iterative algorithms of estimating parameters for the objective functions using the I divergence and IS divergence criterions based on the auxiliary function approach. We have experimentally confirmed that the proposed algorithms outperformed shifted NMFs in the accuracy of source separation. In future, we will examine the effect of setting $M^{(\mathrm{tap})} > 1$ to the source separation accuracy.

6. REFERENCES

[1] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applications Signal Process. Audio Acoust.* IEEE, 2003, pp. 177–180.

[2] D. FitzGerald, M. Cranitch, and E. Coyle, "Shifted non-negative matrix factorisation for sound source separation," in *IEEE/SP 13th Workshop on Statistical Signal Processing*. IEEE, 2005, pp. 1132–1137.

[3] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Mar. 2008, pp. 2069–2072.

[4] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Proc. Int. Conf. Independent Component Analysis and Blind Signal Separation*, 2006, pp. 700–707.

[5] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. Int. Conf. Music Info. Retrieval*, 2007, pp. 381–386.

[6] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Mar. 2008, pp. 109–112.

[7] T. Nakamura, K. Shikata, N. Takamune, and H. Kameoka, "Harmonic-temporal factor decomposition incorporating music prior information for informed monaural source separation," in *Proc. Int. Symposium Music Info. Retrieval*, 2014, pp. 623–628.

[8] "2014: Multiple Fundamental Frequency Estimation & Tracking Results - MIREX Wiki," `http://www.music-ir.org/mirex/wiki/2014:Multiple_Fundamental_Frequency_Estimation_%26_Tracking_Results`.

[9] B. Fuentes, R. Badeau, and G. Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription," *IEEE Trans. Acoust., Speech, and Language Process.*, vol. 21, no. 9, Sept. 2013.

[10] R. Jaiswal, D. FitzGerald, E. Coyle, and S. Rickard, "Towards shifted NMF for improved monaural separation," in *Proc. 24th IET Irish Signals and Systems Conference*, July 2013.

[11] S. Sagayama, H. Kameoka, and T. Nishimoto, "Specmurt analysis: A piano-roll-visualization of polyphonic music signal by deconvolution of log-frequency spectrum," in *Proc. ISCA Tutorial and Research Workshops on Statistical and Perceptual Audition*.

[12] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S Sagayama, "Specmurt analysis of polyphonic music signals," *IEEE Trans. Acoust., Speech, and Language Process.*, vol. 16, no. 3, pp. 639–650, Mar. 2008.

[13] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2007, vol. 1, pp. I–53–I–56.

[14] H. Kameoka and K. Kashino, "Composite autoregressive system for sparse source-filter representation of speech," in *IEEE International Symposium on Circuits and Systems*, 2009, pp. 2477–2480.

[15] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 45–48.

[16] J. J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F. J. Canadas-Quesada, "Musical instrument sound multi-excitation model for non-negative spectrogram," *IEEE J. Sel. Topics. Signal Process.*, vol. 5, no. 6, pp. 1144–1158, Oct. 2011.

[17] H. Kirchhoff, S. Dixon, and A. Klapuri, "Missing template estimation for user-assisted music transcription," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 26–30.

[18] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proc. Int. Conf. Independent Component Analysis and Blind Signal Separation*, 2004, pp. 494–499.

[19] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Number 30. SIAM, 1970.

[20] D. R. Hunter and K. Lange, "Quantile regression via an mm algorithm," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 60–77, 2000.

[21] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds., 2006, pp. 267–296.

[22] H. Kameoka, "Statistical speech spectrum model incorporating all-pole vocal tract model and f_0 contour generating process model," in *IEICE Technical Report*, Nov. 2010, vol. 110, pp. 29–34, in Japanese.

[23] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[24] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE J. Sel. Topics. Signal Process.*, vol. 5, no. 6, pp. 1205–1215, 2011.

[25] H. Kameoka, T. Tahara, T. Nishimoto, and S. Shigeki, "Signal processing method and device," Nov. 2008, Japan Patent JP2008-281898.

[26] T. Nakamura and H. Kameoka, "Fast signal reconstruction from magnitude spectroggram of continuous wavelet transform based on spectrogram consistency," in *Proc. Int. Conf. Digital Audio Effects*, 2014, pp. 129–135.

[27] H. Kameoka, *Statistical Approach to Multipitch Analysis*, Ph.D. thesis, The University of Tokyo, Mar. 2007.

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