Nonnegative Matrix Factorization with Markov-chained Bases for Modeling Time-varying patterns in Music Spectrograms

Masahiro Nakano[†], Jonathan Le Roux[‡], Hirokazu Kameoka[‡], Yu Kitano[†], Nobutaka Ono[†], and Shigeki Sagayama[†]

[†]Graduate School of Information Science and Technology, The University of Tokyo
[‡]NTT Communication Science Laboratories, NTT Corporation

Abstract. *This paper presents a new sparse representation for polyphonic music signals. The goal is to learn the time-varying spectral patterns of musical instruments, such as attack of the piano or vibrato of the violin in polyphonic music signals without any prior information. We model the spectrogram of music signals under the assumption that they are composed of a limited number of components which are composed of Markov-chained spectral patterns. The proposed model is an extension of nonnegative matrix factorization (NMF). An efficient algorithm is derived based on the auxiliary function method.*

Key words: Nonnegative matrix factorization, Sparse signal representation, Source separation, Markov chain, Auxiliary function

1 Introduction

The use of sparse representation in acoustic signal processing has been a very active area of research in recent years, with very effective algorithms based on nonnegative matrix factorization (NMF) [1] and sparse coding. These are typically based on a simple linear model.

NMF, in particular, has been applied extensively with considerable success to various problems including automatic music transcription, monaural sound source separation [2]. NMF is able to project all signals that have the same spectral shape on a single basis, allowing one to represent a variety of phenomena efficiently using a very compact set of spectrum bases. However, because NMF is also fundamentally a dimension reduction technique, a lot of information on the original signal is lost. This is in particular what happens when assuming that the spectrum of the note of a musical instrument can be represented through a single spectral basis whose amplitude is modulated in time, while its variations in time are actually much richer. For example, a piano note would be more accurately characterized by a succession of several spectral patterns such as "attack", "decay", "sustain" and "release". As another example, singing voices and stringed instruments feature a particular musical effect, vibrato, which can

be characterized by its "depth", (the amount of pitch variation), and its "speed", (the speed at which the pitch varies). Learning such time-varying spectra with standard NMF would require to use a large number of bases, and some post-processing to group the bases into single events.

In this paper, we propose a new sparse representation, "NMF with Markov-chained bases" for modeling time-varying patterns in music spectrograms. Our model represents a single event as a succession of spectral patterns. The proposed model is presented in Section 2, together with the derivation of an efficient algorithm to optimize its parameters. We present basic experimental results in Section 3.

2 NMF with Markov-chained bases

2.1 Presentation of the model

Most algorithms for unsupervised sound source separation are based on a signal model where the magnitude or power spectrogram $\mathbf{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\Omega\times T}$, where $\omega = 1,\cdots,\Omega$ is a frequency bin index, and $t = 1,\cdots,T$ is a time frame index, is factorized into nonnegative parameters, $\mathbf{H} = (H_{\omega,d})_{\Omega\times D} \in \mathbb{R}^{\geq 0,\Omega\times D}$ and $\mathbf{U} = (U_{d,t})_{D\times T} \in \mathbb{R}^{\geq 0,D\times T}$. This can be written as

$$Y_{\omega,t} = \sum_{d=1}^{D} H_{\omega,d}U_{d,t} \ , \tag{1}$$

where D is the number of bases $\mathbf{h}_d = [H_{1,d},\cdots,H_{\Omega,d}]$. The term *component* is used to refer to one basis $\mathbf{h}_d$ and its time-varying gain $U_{d,t}$. The bases can be considered as spectral patterns which are frequently observed.

Hopefully, one component should represent a single event. However, the spectrum of instrument sounds is actually in general nonstationary. Each source will thus tend to be modeled as a sum several components, leading to the difficult problem of determining which source each component belongs to. Recently, an extension of NMF where temporal activations become time/frequency activations based on a source/filter model [3] have been proposed to overcome this problem. However, it has not been clarified whether the source/filter model is fit for auditory stream composed of various origin, such as the piano. For example, "attack" caused by a keystroke has energies on wide-band spectrum, while "sustain" has a harmonic structure.

In contrast with the above approaches, we focus on the hierarchical structure of the sounds produced by musical instruments, and model the spectrogram of music signals under the assumption that they are composed of spectral patterns which have themselves a limited number of Markov-chained states. Concretely, we assume that each basis $\mathbf{h}_d$ has Φ states, the transitions between those states being constrained and only one state being activated at each time t.

We attempt to model the spectrogram again based on $\mathbf{H} = (H_{\omega,\phi,d})_{\Omega\times\Phi\times D} \in \mathbb{R}^{\geq 0,\Omega\times\Phi\times D}$ Here, $\mathbf{P} = (P_{\phi,t,d})_{\Phi\times T\times D} \in \mathbb{R}^{\geq 0,\Phi\times T\times D}$ is binary to show which basis is activated at time t, i.e., $P_{\phi,t,d} = 1$ if $\mathbf{h}_d^{(\phi)}$ is activated at time t, and

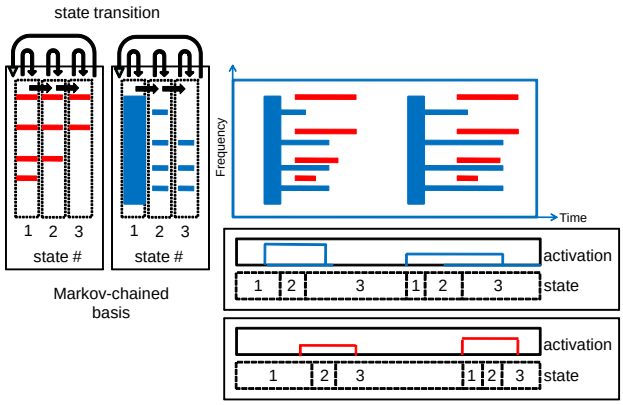


Fig. 1. Diagram of NMF with Markov-chained bases.

$P_{\phi,t,d} = 0$ otherwise. Then, the proposed model can be written as

$$Y_{\omega,t} = \sum_{d,\phi} H_{\omega,\phi,d} P_{\phi,t,d} U_{d,t} \quad . \tag{2}$$

Note that $\mathbf{P}$ does not show the probabilities that state ϕ is activated at time t and the class $\mathfrak{S}$ of all possible $\mathbf{P}$ is defined by the topology of the Markov chain, for example left-to-right model or ergodic model.

2.2 Problem setting

Given an observed spectrogram, we would like to find the optimal estimates of $\mathbf{H}$, $\mathbf{U}$ and $\mathbf{P}$. The standard NMF algorithm proposed by Lee and Seung [1] performs the decomposition by minimizing the reconstruction error between the observation and the model while constraining the matrices to be entry-wise non-negative. Our estimation can also be written as optimization problem, similarly to standard NMF. Various measures for standard NMF have been proposed. We choose here the following β-divergence, which has been widely used,

$$\mathcal{D}_\beta(y|x) = \frac{\left(y^\beta + (\beta-1)x^\beta - \beta y x^{\beta-1}\right)}{\beta(\beta-1)} \quad . \tag{3}$$

Note that the above definition can be properly extended by continuity to $\beta = 0$ and $\beta = 1$ [5]. Let θ denote the set of all parameters $\{(H_{\omega,\phi,d})_{\Omega\times\Phi\times D}, (U_{d,t})_{D\times T}, (P_{\phi,t,d})_{\Phi\times T\times D}\}$. We then need to solve the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad \mathcal{J}(\theta) = \textstyle\sum_{\omega,t} \mathcal{D}_\beta\left(Y_{\omega,t} | \sum_{d,\phi} H_{\omega,\phi,d} P_{\phi,t,d} U_{d,t}\right) \\ &\text{subject to } \forall\omega,\phi,d,\ H_{\omega,\phi,d} \geq 0,\ \forall d,t,\ U_{d,t} \geq 0,\ \mathbf{P} \in \mathfrak{S} \ . \end{aligned} \tag{4}$$

In this paper, we assume that the transition probabilities of the Markov-chained bases are uniform. Thus, the cost of each of the paths is regarded as constant and ignored in Eq. (4). In the following, we will refer to the algorithm to solve NMF with Markov-chained bases as "MNMF".

Note that our model can also be expressed as Factorial Hidden Markov Model for specific βs. The case of $\beta = 0$ reduces to Factorial scaled Hidden Markov Model which reduces to NMF with Itakura-Saito (IS) divergence and Gaussian Scaled Mixture Model [4]. These probabilistic model may achieve the extension of NMF with the Euclidean distance ($\beta = 2$), the generalized Kullback-Leibler divergence ($\beta = 1$) and IS divergence ($\beta = 0$) based on a statistical approach to NMF [5, 6]. However, it is not clarified whether statistical approach can apply to NMF with various measures, such as β-divergence.

2.3 Iterative algorithm

Our derivation is based on a principle called the auxiliary function method, similar to [1]. Let $G(\theta)$ denote an objective function to be minimized w.r.t. a parameter θ. A function $G^+(\theta, \hat{\theta})$ which satisfies $G(\theta) = \min_{\hat{\theta}} G^+(\theta, \hat{\theta})$ is then called an auxiliary function for $G(\theta)$, and $\hat{\theta}$ an auxiliary variable. The function $G(\theta)$ is easily shown to be non-increasing through the following iterative update rules: $\hat{\theta}^{(s+1)} \leftarrow \mathrm{argmin}_{\hat{\theta}} G^+(\theta^{(s)}, \hat{\theta})$ and $\theta^{(s+1)} \leftarrow \mathrm{argmin}_{\theta} G^+(\theta, \hat{\theta}^{(s+1)})$, where $\hat{\theta}^{(s+1)}$ and $\theta^{(s+1)}$ denote the updated values of $\hat{\theta}$ and θ after the s-th step.

An auxiliary function for standard NMF with β-divergence has been proposed [8]. This strategy can apply to our problem. $\hat{\theta}$ denotes auxiliary variables $\{(\lambda_{\omega,t,k})_{\Omega\times T\times K}, (Z_{\omega,t})_{\Omega\times T}\}$ $(\forall k, \lambda_{\omega,t,k} \geq 0, \sum_k \lambda_{\omega,t,k} = 1, Z_{\omega,t} \in \mathbb{R})$ for convenience. We obtain the following auxiliary function:

$$\mathcal{J}^+(\theta, \hat{\theta}) = \sum_{\omega,t} \frac{Y_{\omega,t}}{\beta(\beta-1)} + \sum_{\omega,t} \begin{cases} \mathcal{R}^{(\beta)}_{\omega,t} - Y_{\omega,t} Q^{(\beta-1)}_{\omega,t} & (\beta < 1) \\ \mathcal{Q}^{(\beta)}_{\omega,t} - Y_{\omega,t} Q^{(\beta-1)}_{\omega,t} & (1 \leq \beta \leq 2) \\ \mathcal{Q}^{(\beta)}_{\omega,t} - Y_{\omega,t} R^{(\beta-1)}_{\omega,t} & (\beta > 2) \end{cases} , \quad (5)$$

where $\mathcal{Q}^{(\beta)}_{\omega,t} = (1/\beta) \sum_{d,\phi} \lambda_{\omega,t,d} P_{\phi,t,d} (H_{\omega,\phi,d} U_{d,t} / \lambda_{\omega,t,d})^{\beta}$ and $\mathcal{R}^{(\beta)}_{\omega,t} = (1/\beta) Z^{\beta}_{\omega,t}$ $+ Z^{\beta-1}_{\omega,t} (\sum_{d,\phi} H_{\omega,\phi,d} P_{\phi,t,d} U_{d,t} - Z_{\omega,t})$. $\mathcal{J}^+(\theta, \hat{\theta})$ is minimized w.r.t. $\hat{\theta}$ when

$$\lambda_{\omega,t,d} = \frac{\sum_{\phi} H_{\omega,\phi,d} P_{\phi,t,d} U_{d,t}}{\sum_{\phi,d} H_{\omega,\phi,d} P_{\phi,t,d} U_{d,t}} \ , \ Z_{\omega,t} = \sum_{\phi,d} H_{\omega,\phi,d} P_{\phi,t,d} U_{d,t} \ . \quad (6)$$

Minimizing $\mathcal{J}^+(\theta, \hat{\theta})$ w.r.t. $\mathbf{P} \in \mathfrak{S}$ is a search problem

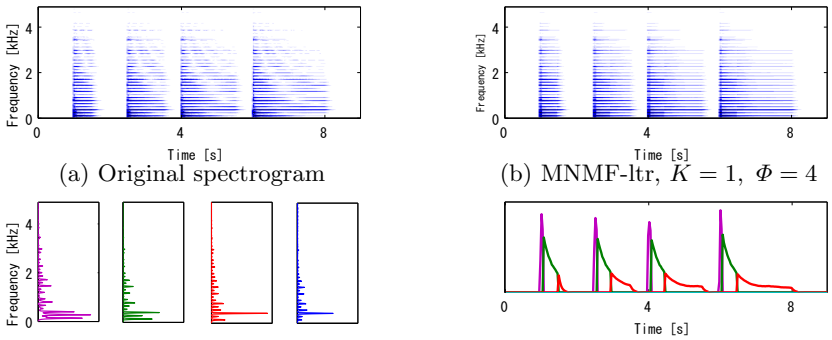
$$\mathbf{P} \leftarrow \underset{\mathbf{P}\in\mathfrak{S}}{\mathrm{argmin}} \left\{ \mathcal{J}^+(\theta, \hat{\theta}) \right\} \ , \quad (7)$$

which can be straightforwardly solved using the Viterbi algorithm. Differentiating $\mathcal{J}^+(\theta, \hat{\theta})$ partially w.r.t. $H_{\omega,\phi,t}$ and $U_{d,t}$, and setting to zero, we obtain update rules for $H_{\omega,\phi,d}$ and $U_{d,t}$:

$$H_{\omega,\phi,d} \leftarrow \left(\frac{\sum_t \alpha^{\beta-2}_{\omega,t} Y_{\omega,t} \gamma^{\mathrm{Max}\{2-\beta,\ 0\}}_{\omega,t,d} P_{\phi,t,d} U_{d,t}}{\sum_t \alpha^{\beta-1}_{\omega,t} \gamma^{\mathrm{Min}\{1-\beta,\ 0\}}_{\omega,t,d} P_{\phi,t,d} U_{d,t}} \right)^{\varphi(\beta)} \ , \quad (8)$$

$$U_{d,t} \leftarrow U_{d,t} \left(\frac{\sum_{\omega} \alpha^{\beta-2}_{\omega,t} Y_{\omega,t} \gamma_{\omega,t,d}}{\sum_{\omega} \alpha^{\beta-1}_{\omega,t} \gamma_{\omega,t,d}} \right)^{\varphi(\beta)} \ , \quad (9)$$

where $\alpha_{\omega,t} = \sum_{d,\phi} H_{\omega,\phi,d} P_{\phi,t,d} U_{d,t}$, $\gamma_{\omega,t,d} = \sum_{\phi} H_{\omega,\phi,d} P_{\phi,t,d}$ and $\varphi(\beta) = 1/(2-\beta)\ (\beta < 1),\ 1\ (1 \leq \beta \leq 2),\ 1/(\beta-1)\ (\beta > 2)$.



(a) Original spectrogram (b) MNMF-ltr, $K = 1,\ \Phi = 4$

(c) Markov-chained bases, $K = 1,\ \Phi = 4$ (d) Activation, $K = 1,\ \Phi = 4$

Fig. 2. Original spectrogram of the extract of the piano (MIDI) (a), reconstructed spectrograms (b), the bases (c) and the activation (d) learned by MNMF. MNMF was able to decompose the evolution of the spectrum as a succession of several part, "attack", "sustain", "decay" and "release".

2.4 Update scheduling

The algorithm described above converges quickly. However, it often falls into unexpected stationary points. **H** and **U** at the early stage of iterations are not useful for estimating **P**. **P** fixed using unrealistic **H** and **U** induces in turn **H** and **U** in unexpected directions. We thus improve the update rule for **P** by introducing an updating schedule. Here, let the scheduling parameter, $k_\phi^{(s)}$, at s-th iteration satisfy $\forall\phi,\ k_\phi \geq 0,\ \sum_\phi k_\phi = 1,\ k_1^{(s+1)} \geq k_1^{(s)},\ k_1^{(S)} = 1,\ k_\phi^{(s+1)} \leq k_\phi^{(s)},\ k_\phi^{(S)} = 0\ (\phi = 2, \cdots, \Phi)$. We replace the update rule, Eq. (7) by

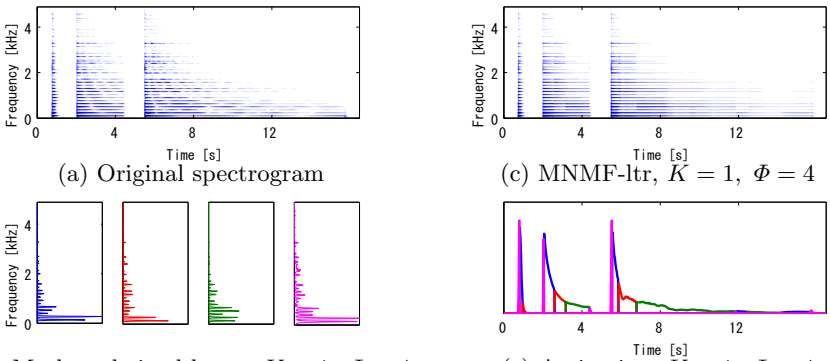
$$P_{\phi,t,d}^{(s+1)} \leftarrow \sum_{n=1}^{\Phi} k_n^{(s+1)} \hat{P}_{\phi-(n-1),t,d} \ , \tag{10}$$

where $\forall n,\ \hat{P}_{-(n-1),t,d} = \hat{P}_{\Phi-(n-1),t,d}$ and $(\hat{P}_{\phi,t,d})_{\Phi\times T\times D} = \mathrm{argmin}_{\mathbf{P}\in\mathfrak{S}}\, \mathcal{J}^+(\theta, \hat{\theta})$. As a result, at each iteration the auxiliary function is not minimized anymore. However, convergence is guaranteed.

3 Simulation results

In this section, some results on the application of our algorithm to audio signals. All data were downmixed to mono and downsampled to 16kHz. The magnitude spectrogram was computed using the short time Fourier transform with 32 ms long Hanning window and with 16 ms overlap. The state transitions of the bases in MNMF were modeled using left-to-right (-ltr) and ergodic (-erg) models.

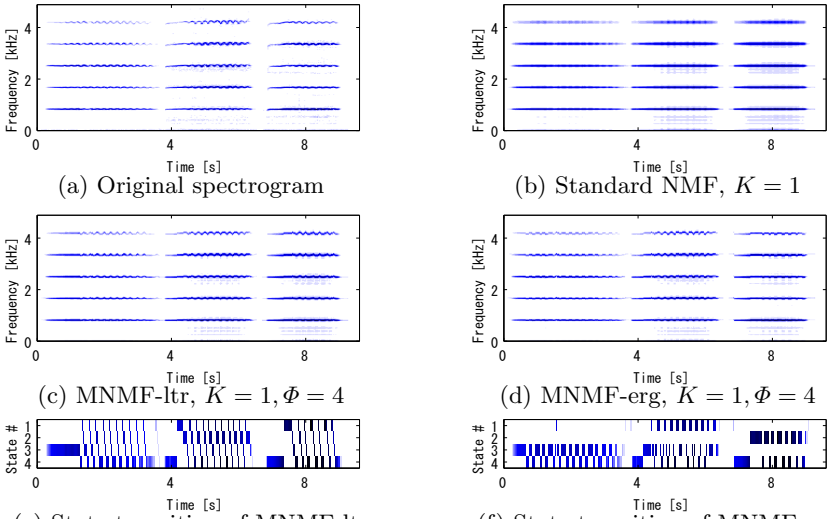
At first, we tested whether the algorithm was able to learn in an unsupervised way the time-varying spectral patterns from notes with a unique pitch. The proposed method was applied to a piano note (C3) synthesized from MIDI, a piano note (C3) recorded from RWC-MDB-I-2001 No.1 [7] and a violin note (A♭) recorded from RWC-MDB-I-2001 No.15. We used MNMF with $\beta = 1$ (the



(a) Original spectrogram (c) MNMF-ltr, $K = 1,\ \Phi = 4$

(d) Markov-chained bases, $K = 1,\ \Phi = 4$ (e) Activation, $K = 1,\ \Phi = 4$

Fig. 3. Original spectrogram of the extract of the piano (RWC database) (a), reconstructed spectrograms (b) and (d), the bases (c) and the activation (d) learned by MNMF. Time-varying spectral patterns are also learned.



(a) Original spectrogram (b) Standard NMF, $K = 1$

(c) MNMF-ltr, $K = 1, \Phi = 4$ (d) MNMF-erg, $K = 1, \Phi = 4$

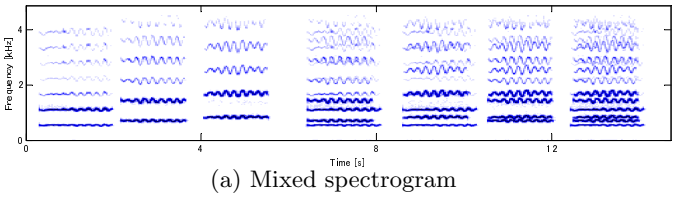
(e) State transition of MNMF-ltr (f) State transition of MNMF-erg

Fig. 4. Original spectrogram of the extract of the violin (RWC database) (a), reconstructed spectrograms (b), (c) and (d), and the activation (e) and (f) learned by MNMF. The topology of the Markov chain affects the state transitions.

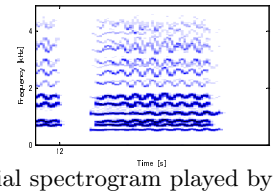
generalized Kullback-Leibler divergence). As shown in Fig. 2, 3 and 4, time-varying spectral patterns are learned in an unsupervised way.

Next, we applied our model to a mixture of vocal signals taken from RWC-MDB-I-2001 No.45. The sequence is composed of 3 notes (D♭, F, A♭): first, each note is played alone in turn, then all combination of two notes are played and finally all notes are played simultaneously. The result is shown in Fig. 5.

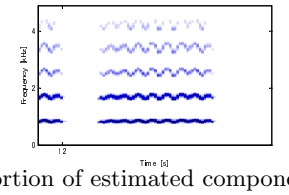
Finally, our model was applied to sound source separation. The tested signals are RWC signal (as shown in Fig. 5) and audio data recorded in real-



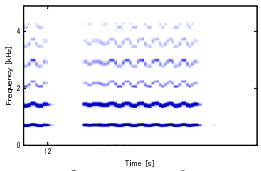
(a) Mixed spectrogram



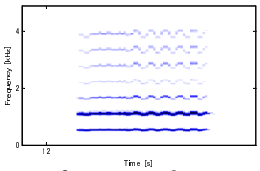
(b) Partial spectrogram played by all notes.



(c) Portion of estimated component: A♭



(d) Portion of estimated component: F



(e) Portion of estimated component: D♭

Fig. 5. Mixed spectrogram (a) and (b), and one portion of estimated spectrogram of each component (c), (d) and (e).

Table 1. Source Separation Performance

Algorithms	RWC signal		Real-world audio signal	
	Number of **H** and **U**	SNR(dB)	Number of **H** and **U**	SNR(dB)
NMF($D = 3$)	2919	3.55	2871	6.49
NMF($D = 6$)	5838	7.11	5742	9.61
NMF($D = 9$)	**8757**	**10.63**	**8613**	**10.50**
NMF($D = 12$)	11676	13.51	11484	11.70
MNMF-erg($\Phi = 4$)	7536	9.62	7488	9.57
MNMF-erg($\Phi = 5$)	**9075**	**11.02**	**9027**	**9.84**

world conditions from male vocal in the room size of 5.5m × 3.5m × 3m by IC recorder, whose sequence is composed similar to RWC signal (A, D♭ and E instead of D♭, F and A♭). We use the signal-to-noise ratio (SNR) between each component and source as the measure for assigning components to sources. The measure was calculated between the magnitude spectrograms $I^{(m)}_{\omega,t}$ and $\hat{I}^{(n)}_{\omega,t}$ of the mth reference and the nth separated component, respectively, $\mathrm{SNR} = 10\log_{10}(\sum_{\omega,t}(I^{(m)}_{\omega,t})^2/\sum_{\omega,t}(I^{(m)}_{\omega,t} - \hat{I}^{(n)}_{\omega,t})^2)$. The SNR was averaged over all the sources to get the separation performance and each algorithm was run 10 times with random initializations. We set the number of bases to $D = 3$ for MNMF. This means that one component with Φ spectral patterns may be expected to represent one source having the perceveid pitch of a note with vibrato. Standard NMF was used as the baseline. NMF with $D = 3$ was expected to separate each source into a single component, while for NMF with $D > 3$

each source was expected to be split into the sum of several components. A component n is assigned to the source m which leads to the highest SNR. As reported by [2], using NMF with a large number of components and clustering them using the original signals as references may produce unrealistically good results. One should thus keep in mind when comparing the results of our method with those of NMF that the experimental conditions were strongly biased in favor of NMF. As shown in Table 1, we can see that MNMF performs as well as standard NMF when the total number of parameters $\mathbf{H}$ and $\mathbf{H}$ is similar ($\mathbf{P}$ is excluded), although again for NMF the original signals need to be used to cluster the extracted components.

4 Concluding remarks

We developed a new framework for the sparse representation of audio signals. The proposed model is extension of NMF, in which the bases consist of state transition. We derived an efficient algorithm for the optimization of the model based on the auxiliary function method. Future work will includes the extension of our model to automatic estimation of the number of bases and states.

Acknowledgements

This research was partially supported by CrestMuse Project under JST from Mext, Japan, and Grant-in-Aid for Scientific Research (KAKENHI) (A) 20240017.

References

1. D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, vol. 13, pp. 556–562, Dec. 2001.
2. T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1066–1074, Mar. 2007.
3. R. Hennequin, R. Badeau, and B. David, "NMF with time-frequency activations to model non stationary audio events," in *Proc. ICASSP*, pp. 445–448, Mar. 2010.
4. A. Ozerov, C. Fevotte and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. WASPAA*, 2009.
5. 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, pp. 793–830, Mar. 2009.
6. C. Févotte and A. T. Cemgil, "Nonnegative matrix factorizations as probabilistic inference in composite models," in *Proc. EUSIPCO*, vol. 47, pp. 1913–1917, 2009.
7. M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in *Proc. ISMIR*, pp. 287–288, 2002.
8. M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Proc. MLSP*, 2010.