BAYESIAN NONPARAMETRIC SPECTROGRAM MODELING BASED ON INFINITE FACTORIAL INFINITE HIDDEN MARKOV MODEL

Masahiro Nakano[†]*, Jonathan Le Roux*[‡]*, Hirokazu Kameoka*[‡]*, Tomohiko Nakamura*[†]*, Nobutaka Ono*[†] *and Shigeki Sagayama*[†]

[†]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, NTT Corporation,
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan

ABSTRACT

This paper presents a Bayesian nonparametric latent source discovery method for music signal analysis. In audio signal analysis, an important goal is to decompose music signals into individual notes, with applications such as music transcription, source separation or note-level manipulation. Recently, the use of latent variable decompositions, especially nonnegative matrix factorization (NMF), has been a very active area of research. These methods are facing two, mutually dependent, problems: first, instrument sounds often exhibit time-varying spectra, and grasping this time-varying nature is an important factor to characterize the diversity of each instrument; moreover, in many cases we do not know in advance the number of sources and which instruments are played. Conventional decompositions generally fail to cope with these issues as they suffer from the difficulties of automatically determining the number of sources and automatically grouping spectra into single events. We address both these problems by developing a Bayesian nonparametric fusion of NMF and hidden Markov model (HMM). Our model decomposes music spectrograms in an automatically estimated number of components, each of which consisting in an HMM whose number of states is also automatically estimated from the data.

Index Terms— Nonnegative matrix factorization (NMF), Hierarchical Dirichlet process (HDP), Infinite hidden Markov model (iHMM), Gamma process, Collapsed variational Bayes (CVB)

1. INTRODUCTION

In acoustic signal processing, the use of latent variable decompositions, especially NMF [1], has been a very active area of research. In the context of audio signal analysis based on NMF, the observed signal (magnitude or power spectrogram) is approximated by a linear combination of spectral bases which can be regarded as the representatives of the various spectral patterns in the signal.

This paper focuses on the two problems which standard NMF faces in music signal analysis. Firstly, one needs to deal with the non-stationarity of instrument sounds [2, 3]. Hopefully, each spectral basis should be the representative of a single instrument sound. In standard NMF, this implies that the elementary components (which one would like to correspond to one pitch of one instrument) of the analyzed sound are assumed to be nearly stationary. However, real world sounds often exhibit non-stationary spectral characteristics. 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. Spectra of instrument sounds with such variations are typically difficult to handle by standard NMF. Indeed, a single instrument sound will tend to be modeled as the sum of several spectral bases, leading to the difficult problem of determining which sources each component belongs to. Standard NMF thus requires some post-processing to group the bases into single notes. Secondly, it is difficult to estimate the number of notes which are included in the observed polyphonic music signals [4, 5]. Most methods require that the number of sources be specified in advance, or found with model selection techniques [5]. This problem is related to the non-stationarity of the spectrum of instrumental sounds, because the variations in the spectral patterns of a sound make it even more difficult to estimate the number of sources. These problems should thus be addressed simultaneously.

We do so by developing a Bayesian nonparametric fusion of NMF and hidden Markov model (HMM), which can be regarded as an extension of the factorial hidden Markov model (FHMM). The number of components and the number of states have a strong impact on the standard FHMM. To overcome this issue, we place here a prior over all their possible combinations via a Bayesian nonparametric framework. Our model decomposes music spectrograms in an automatically estimated number of components, each of which consisting in an HMM whose number of states is also automatically estimated from the data.

Conventional Bayesian nonparametric NMF, here reformulated with the generalized KL-divergence, is explained in section 2. Section 3 presents our model based on Bayesian nonparametric combined NMF and HMM. An efficient inference algorithm based on variational Bayes (and collapsed VB) is derived in Section 4. Section 5 discusses potential improvements to the model based on a different construction of the prior. Experimental results are shown in Section 6.

2. GAMMA PROCESS NONNEGATIVE MATRIX FACTORIZATION

An NMF approach to audio signal analysis is typically based on the assumption that a magnitude or power spectrogram $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\Omega\times T}$, where $\omega = 1,\cdots,\Omega$ is a frequency bin index, and $t = 1,\cdots,T$ is a time frame index, can be modeled as the product of two non-negative matrices, $\boldsymbol{H} = (H_{\omega,d})_{\Omega\times D} \in \mathbb{R}^{\geq 0,\Omega\times D}$ and $\boldsymbol{U} = (U_{d,t})_{D\times T} \in \mathbb{R}^{\geq 0,D\times T}$. This can be written as $Y_{\omega,t} \approx \sum_d H_{\omega,d}U_{d,t}$, where D is the number of bases $\mathbf{h}_d = [H_{1,d},\cdots,H_{\Omega,d}]^T$. The term "component" is often used to refer to a spectral portion corresponding to $\mathbf{h}_d$ and $U_{d,t}$. Although the discrepancy measure between $\boldsymbol{Y}$ and $\boldsymbol{HU}$ can be defined in

978-1-4577-0693-6/11/$26.00

many ways, we will focus in this paper on the use of the generalized Kullback-Leibler (KL) divergence in the magnitude domain, since this choice has been shown to lead to good music source separation performance [6].

The standard NMF assumes that the number of components D is known in advance. In practice, we rarely know the adequate value for this number and must resort to expensive techniques such as model selection to estimate it. Recently, Bayesian nonparametric approaches have proven to be effective for the problem of automatically inferring model complexities [5, 7]. The Gamma process nonnegative matrix factorization (GaP-NMF) [5] introduces a hidden vector of nonnegative values $\boldsymbol{\theta}$, where each element θ_d is the overall gain of the corresponding source d. While the original GaP-NMF [5] is basically a nonparametric extension of the Itakura-Saito divergence-based NMF, we start here by formulating a nonparametric extension of the generalized KL divergence-based NMF using the Gamma process prior. The observed spectrogram $\boldsymbol{Y}$ and the overall gains $\boldsymbol{\theta}$ are expressed according to the following generative process: $C_{\omega,t,d} \sim \mathrm{Poisson}(\theta_d H_{\omega,d} U_{d,t})$, $\theta_d \sim \mathrm{Gamma}(\eta/D, \eta\lambda)$, $Y_{\omega,t} \sim \delta(Y_{\omega,t} - \sum_d C_{\omega,t,d})$ where $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_D$ can be interpreted as latent components. Note that D works as the truncation level. As D increases towards infinity, the overall gains $\boldsymbol{\theta}$ can be considered to be drawn from a Gamma process with shape parameter η and inverse-scale parameter $\eta\lambda$ [8, 5]. When the truncated level D is sufficiently large, we can expect that an adequate number of components will be active and the rest will be suppressed. We thus don't need to determine the adequate number of components in advance. Instead, we only have to prepare a sufficiently large number D of them.

3. BAYESIAN NONPARAMETRIC COMBINED NMF AND HMM FOR MODELING MUSIC SPECTROGRAMS

GaP-NMF gives us an adequate number of components required by the model to describe the data, but we would like each of these components to correspond to a single event, such as "C♯ of the piano" or "A♭ of the violin". This would likely be the case if we could assume that the spectrum of a note of a musical instrument can be represented through a single spectral basis whose amplitude is modulated in time, but its variations in time are actually much richer.

Recently, an HMM-based approach has been proposed to overcome this problem [2]. The spectrogram of the observed signal is modeled under the assumption that it is composed of spectral patterns which are themselves composed of a limited number of Markov-chained states. To introduce this concept into GaP-NMF, we consider state-transition bases $\boldsymbol{H} = \{(H^{(k)}_{\omega,1})_{\Omega\times K}, \ldots, (H^{(k)}_{\omega,D})_{\Omega\times K}\}$, where $(H^{(k)}_{\omega,d})_{\Omega}$ denotes the k-th possible state for the spectral basis of the d-th component and K denotes the number of states. If we let $\boldsymbol{Z} = (Z_{d,t})_{D\times T} \in \mathbb{N}$ denote which spectral basis state of the d-th component is activated at time t, the generative model can be expressed as $H^{(k)}_{\omega,d} \sim \mathrm{Gamma}(a_H, b_H)$, $U_{d,t} \mid W_{d,t} \sim \mathrm{Gamma}(a_U, a_U W_{d,t})$, $W_{d,t} \mid U_{d,t-1} \sim \mathrm{Gamma}(a_U, a_U U_{d,t-1})$, $\theta_d \sim \mathrm{Gamma}(\eta/D, \eta\lambda)$ and

$$C_{\omega,t,d} \mid (H^{(k)}_{\omega,d})_{k=1}^{K}, U_{d,t}, Z_{d,t} \sim \mathrm{Poisson}\left(\theta_d H^{(Z_{d,t})}_{\omega,d} U_{d,t}\right) \ ,$$
$$Y_{\omega,t} \sim \delta\Big(Y_{\omega,t} - \sum_d C_{\omega,t,d}\Big) \ , \qquad (1)$$

where $\boldsymbol{W} = (W_{d,t})_{D\times T}$ are auxiliary variables. The Gamma chain of $\boldsymbol{U}$ and $\boldsymbol{W}$ promotes temporal continuity, often encountered in real-world sounds [13].

The number of states is likely to be an important factor to characterize the diversity of each instrument. Hopefully, the adequate number of states should be assigned in response to the tested instrumental sounds. To address this problem, we introduce a Bayesian nonparametric approach to the state-transition bases, similarly to the infinite HMM [9]. Our model is constructed based on the Bayesian nonparametric HMM using hierarchical Dirichlet process (HDP) [9, 10, 11]. A two-level HDP can be used to develop an HMM with an infinite state space. It is a collection of DPs, $\{G_1, G_2, \ldots\}$, sharing a base distribution G_0: $G_j \sim \mathrm{DP}(\alpha, G_0)$, for each j. G_0 is also drawn from a DP with a base distribution F: $G_0 \sim \mathrm{DP}(\gamma, F)$. α, γ are referred to as concentration parameters, which can be interpreted as controlling the impact of the base distribution. The intuitive interpretation of the HDP in our model can briefly be described as follows: the base distribution F is a distribution over the nonnegative spectrum space and draws independently the atoms of the top-level DP G_0. As a draw from DP, G_0 is almost surely discrete, leading to the discreteness of the HMM. The fact that all G_j share G_0 enables them to share the same set of atoms with only different weights. It ensures that transitions are allowed only on the same set of atoms.

Various constructions of the HDP-HMM prior have been proposed [9, 12, 11, 15]. Although, as discussed in Section 5, some alternatives may lead to potentially better models, we choose to present in this paper an intuitive construction based on the definition of the HDP and the stick-breaking process [9, 12]. Let $\boldsymbol{\beta}_d = \beta_{d,1}, \beta_{d,2}, \ldots, \beta_{d,k}, \ldots$ be generated from a stick-breaking process: $\boldsymbol{\beta}_d \sim \mathrm{GEM}(\gamma)$, that is, $\beta'_{d,m}|\gamma \sim \mathrm{Beta}(1, \gamma)$, $\beta_{d,m} = \beta'_{d,m} \prod_{l=1}^{m-1}(1 - \beta'_{d,l})$. Let $\boldsymbol{\pi}_{d,k} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}_d)$ where $\boldsymbol{\pi}_{d,k}$ denotes the state-specific transition distribution, that is, $\pi_{d,k,k'}$ shows the state-transition probability from state k to state k'. Then, the states are encouraged to be similar because $\mathbb{E}[\pi_{d,k,k'}|\boldsymbol{\beta}_d] = \beta_{d,k'}$. The whole model can be expressed by combining Eq. (1) and the following generative process:

$$\boldsymbol{\beta}_d \sim \mathrm{GEM}(\gamma) \ , \ \boldsymbol{\pi}_{d,k} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}_d)$$
$$Z_{d,t} \mid Z_{d,t-1}, (\pi_{d,k})_{k=1}^{\infty} \sim \boldsymbol{\pi}_{d,Z_{d,t-1}} \qquad (2)$$

The key property is that only meaningful elements of the unlimited number of states are active in essence. We thus do not need to find the suitable number of states in advance, as it is assigned in response to the tested instrumental sounds.

To summarize, we introduced two infinite models into the standard FHMM. We shall refer to this model as the infinite factorial infinite hidden Markov model (iFiHMM).

4. VARIATIONAL INFERENCE

Inference algorithms for models with an HDP prior are mostly based on sampling methods. For HDP-HMM, various types of Markov chain Monte Carlo methods have been proposed [9, 10, 11]. In the context of Bayesian inference, there are alternatives based on variational Bayes (VB). Especially, for large-scale problems, VB may prove to be one of the most convenient methods. In this paper, we use variational inference for our model because music spectrograms often comprise a large number of parameters.

The VB approach in general assumes a factorized form of the posterior distribution. This implies that the parameters are assumed to be independent of each other, and such an assumption is likely to often degrade the performance of the inference when it is not met. It is the case in iFiHMM, where $\boldsymbol{\pi}_d$ is likely to have a strong impact on $\boldsymbol{Z}_d$. In order to better approximate the posterior, we thus use here

the recently developed collapsed VB (CVB) for HDP, integrating out certain parameters while assuming that other latent variables are independent [12].

To derive collapsed variational updates, we first integrate out $\boldsymbol{\pi}$, leading to a joint distribution over $\boldsymbol{Z}, \alpha, \boldsymbol{\beta}$ as follows:

$$p(\boldsymbol{Z} \mid \alpha, \boldsymbol{\beta}) = \prod_{d,j} \Big\{ \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{d,j,\cdot})} \prod_k \frac{\Gamma(\alpha\beta_{d,k} + n_{d,j,k})}{\Gamma(\alpha\beta_{d,k})} \Big\} \quad (3)$$

where $n_{d,j,k}$ counts the number of times that the state transition from state j to state k is drawn and $n_{d,j,\cdot} = \sum_k n_{d,j,k}$. The gamma function in Eq. (3) makes the hyperparameter posterior updates cumbersome. We thus introduce auxiliary variables $s_{d,j,k}$ taking integral values, similarly to [12]. The joint distribution over the expanded system is obtained as:

$$p(\boldsymbol{Z}, \boldsymbol{s} \mid \alpha, \boldsymbol{\beta}) = \prod_{d,j} \Big\{ \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{d,j,\cdot})} \prod_k \begin{bmatrix} n_{d,j,k} \\ s_{d,j,k} \end{bmatrix} (\alpha\beta_{d,k})^{s_{d,j,k}} \Big\}$$

where $\begin{bmatrix} \cdot \\ \cdot \end{bmatrix}$ denotes the unsigned Stirling numbers of the first kind.

Given the observed spectrogram $\boldsymbol{Y}$, we want to compute the posterior distribution $p(\boldsymbol{C}, \boldsymbol{\theta}, \boldsymbol{H}, \boldsymbol{U}, \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{s} | \alpha, \eta, \lambda, a, b)$. In the following, we will refer to the parameters of interest as Ξ. CVB approximates the posterior with the variational distribution $q(\Xi) = q(\boldsymbol{C})q(\boldsymbol{\theta})q(\boldsymbol{H})q(\boldsymbol{U})q(\boldsymbol{Z})q(\boldsymbol{\beta})q(\boldsymbol{s}|\boldsymbol{Z})$. We optimize the factorized distribution of each parameter iteratively while holding the others fixed. Space constraints allow only for a sketch of the update rules for $q(\boldsymbol{Z})$:

$$\begin{aligned} q(Z_{d,t} = k) \propto \exp\Big(& - \sum_\omega \mathbb{E}[\theta_d H^{(k)}_{\omega,d} U_{d,t}] \\ & + \sum_\omega \mathbb{E}[C_{\omega,t,d} \log H^{(k)}_{\omega,d} U_{d,t}] + \mathbb{E}[\log(\alpha\beta_{d,k} + n^{-t}_{d,z_{d,t-1},k})] \\ & + \mathbb{E}[\log(\alpha\beta_{d,z_{d,t+1}} + n^{-t}_{d,k,z_{d,t+1}} + \mathbb{I}(z_{d,t-1} = k)\mathbb{I}(z_{d,t+1} = k))] \\ & - \mathbb{E}[\log(\alpha + n^{-t}_{d,k,\cdot} + \mathbb{I}(z_{d,t-1} = k))] \Big) \end{aligned}$$

where $n^{-t}_{d,j,k}$ counts the number of times that state transition from state j to state k is drawn, excluding the number of transitions $Z_{d,t-1} \to Z_{d,t}$ or $Z_{d,t} \to Z_{d,t+1}$.

The update rules for $q(\boldsymbol{C})$, $q(\boldsymbol{H})$ and $q(\boldsymbol{U})$ are derived by making a slight modification (adding $\boldsymbol{\theta}$) to [3]. The updates of $q(\boldsymbol{\beta})$ and $q(\boldsymbol{s}|\boldsymbol{Z})$ are derived similarly to [12]. The updates of $q(\theta)$ derived in [5] are generalized to account for the state transitions between spectral bases. It is computationally difficult to obtain some of the expectations. However, we can use the same tractable approximations employed in [12].

5. ALTERNATE CONSTRUCTION FOR THE HDP-HMM PRIOR

It is worth noting that a different VB approach for HDP has recently been introduced in [11, 15], based on the Chinese restaurant franchise (CRF) [9]. The whole model can be expressed by combining Eq. (1) and:

$$\begin{aligned} \boldsymbol{\beta}_d \sim \mathrm{GEM}(\gamma) \ , \ \boldsymbol{\pi}_{d,j} \sim \mathrm{GEM}(\alpha) \ , \ \varsigma_{d,j,i} \sim \boldsymbol{\beta}_d \\ \eta_{d,j,t} \sim \boldsymbol{\pi}_{d,j} \ , \ Z_{d,t} = \varsigma_{d,Z_{d,t-1},\eta_{d,Z_{d,t-1},t}} \end{aligned} \quad (4)$$

Because this construction does not involve variables having a strong influence on others, there is no special incentive to using collapsed VB. Moreover, owing to the full conjugacy of the construction, VB can be straightforwardly applied and the update rules obtained in closed form. This construction thus seems very promising, and we shall report on it in more details in a future publication.

6. EXPERIMENTS

We present some results on the application of our algorithm to audio signals, for fully unsupervised sound separation. All data were downmixed to mono and downsampled to 16 kHz. The magnitude spectrogram was computed using the short time Fourier transform with 32 ms long Hanning window and with 16 ms overlap.

We first generated synthetic data (shown in Fig. 1 (a)) consisting in a mixture of piano (C), violin (E) and flute (G), chosen to have overlapping harmonic components: first, each note is played alone in turn, then all the combinations of two notes are played and finally all notes are played simultaneously. To clarify the effect of the present method, we used as a comparison standard NMF where the number D of bases is fixed to 3. The result is shown in Fig. 1. Standard NMF, even when given the proper D, is unable to represent correctly the time-varying spectra, such as the attack part of the piano and the vibrato of the violin. It would need more spectral bases to capture the time-varying spectra, but some post-processing would then be required to group the bases into single notes. By contrast, the proposed method (truncation levels: $D = 10$ and $K = 30$) automatically factorizes the observed spectrogram into the mainly active components (shown in Fig. 2). The relative level of the remaining components is -109.2 dB. The proposed method is thus able to automatically find the adequate number of components and group the proper number of Markov-chained spectral bases to represent non-stationary spectra.

As an example of real application, we also present results on the task of note-level manipulation of real-world music signals. Some audio samples as well as a description of the procedure we used are available at http://hil.t.u-tokyo.ac.jp/~mnakano/.

7. CONCLUSION

This paper presented a Bayesian nonparametric combined NMF and HMM framework for modeling music spectrograms. We showed through experiments that our model is capable of automatically determining the number of sources and of modeling the variety of the time-varying spectra of each instrument sound, via a Bayesian nonparametric approach. In the future, we will apply this model to various tasks related to music signal analysis. We also plan to extend our model to a semi-supervised setup.

8. REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, pp. 788–791, Oct. 1999.

[2] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. WASPAA*, 2009.

[3] M. Nakano, J. Le Roux, H. Kameoka, N. Ono and S. Sagayama, "Infinite-state spectrum model for music signal analysis," in *Proc. ICASSP*, pp. 1972–1975, 2011.

[4] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization," in *Proc. SPARS*, 2009.

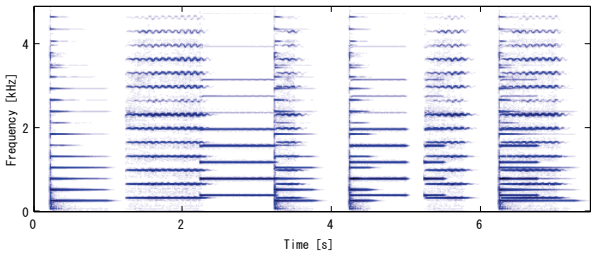
[5] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. ICML*, pp. 641–648, 2010.

[6] D. FitzGerald, M. Cranitch and E. Coyle, "On the use of the Beta divergence for musical source separation," in *Proc. ISSC*, 2009.

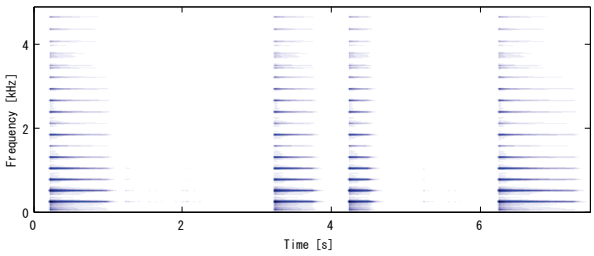
[7] M. N. Schmidt and M. Mørup, "Infinite non-negative matrix factorization," in *Proc. EUSIPCO*, 2010.

[8] J. F. C. Kingman, "Poisson processes," *Oxford University Press*, 1993.

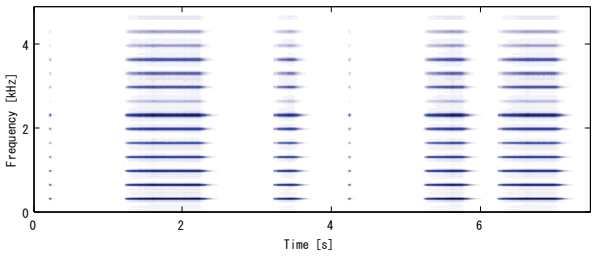
[9] Y. Teh, M. Jordan, M. Beal and D. Blei, "Hierarchical Dirichlet processes," in *Proc. Journal of the American Statistical Association, 101*, pp. 1566–1581, 2004.



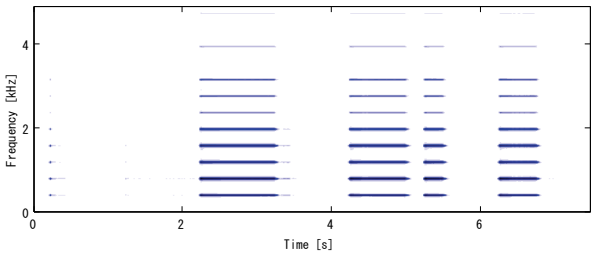
(a) Input spectrogram, a mixture of piano, violin and flute



(b) Estimated model: Piano (C)

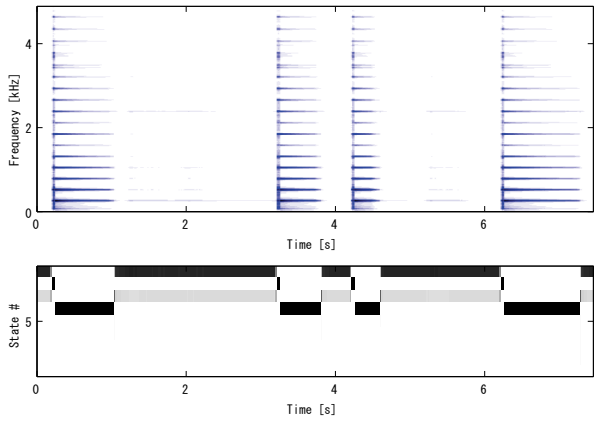


(c) Estimated model: Violin (E)

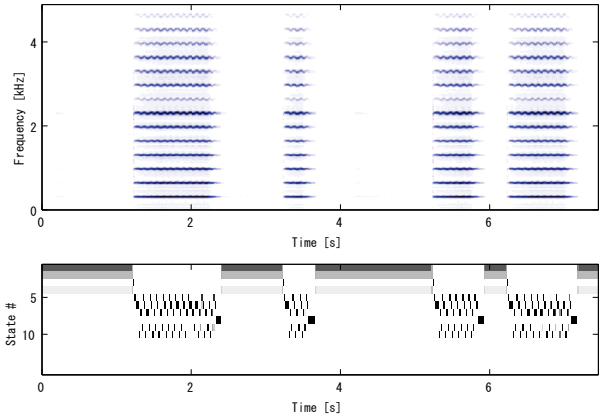


(d) Estimated model: Flute (G)

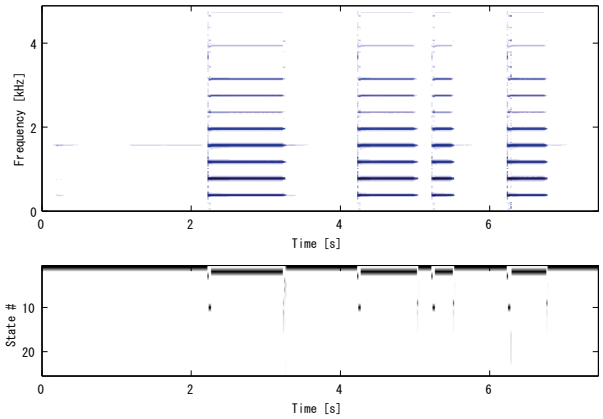
Figure 1: Original spectrogram (a), estimated model $H_{\omega,d}U_{d,t}$ (b), (c) and (d) obtained by the standard NMF (D=3). The standard NMF cannot explicitly represent the time-varying nature of the spectra of instrument sounds. We can see that it does not capture the attack part of piano and flute, and the vibrato of violin.



(a) Estimated model: Piano (C)



(b) Estimated model: Violin (E)



(c) Estimated model: Flute (G)

Figure 2: Estimated model obtained by applying the proposed method to the input signal (Fig. 1 (a)): $\mathbb{E}[\theta_d H_{\omega,d}^{(Z_{d,t})} U_{d,t}]$ (top) and $q(Z_{d,t} = k)$ (bottom) of each component. We set the truncation levels to $D = 10$ and $K = 30$, and hyperparameters to $\alpha = \gamma = 1$, $a_H = b_H = 0.001$, $a_U = 1$, $\eta = 0.1$ and $\lambda = \Omega T / \sum_{\omega,t} Y_{\omega,t}$. Three components are mainly active and the relative level of the remaining components is -109.2 dB. The adequate numbers of components and states are automatically optimized depending on the observed signal.

[10] J. V. Gael, Y. Saatci, Y. W. Teh and Z. Ghahramani, "Beam sampling for the infinite hidden Markov model," in *Proc. ICML*, 2008.

[11] E. B. Fox, E. B. Sudderth, M. I. Jordan and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. ICML*, 2008.

[12] Y. W. Teh, K. Kurihara and M. Welling, "Collapsed variational inference for HDP," in *Proc. NIPS*, 2008.

[13] A. T. Cemgil and O. Dikmen, "Conjugate Gamma Markov random fields for modelling nonstationary sources," in *Proc. ICA*, pp. 697–705, 2007.

[14] 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.

[15] C. Wang, J. Paisley and D. M. Blei, "Online variational inference for the hierarchical Dirichlet process," in *Proc. AISTATS*, 2011.