CONSTRAINED AND REGULARIZED VARIANTS OF NON-NEGATIVE MATRIX FACTORIZATION INCORPORATING MUSIC-SPECIFIC CONSTRAINTS

Hirokazu Kameoka[1,2], *Masahiro Nakano*[2], *Kazuki Ochiai*[1], *Yutaka Imoto*[1], *Kunio Kashino*[2], *Shigeki Sagayama*[1]

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

ABSTRACT

Music spectrograms typically have many structural regularities that can be exploited to help solve the problem of decomposing a given spectrogram into distinct musically meaningful components. In this paper, we introduce new variants of the non-negative matrix factorization concept that incorporate music-specific constraints.

Index Terms— Non-negative Matrix Factorization, Music transcription, Source separation

1. INTRODUCTION

Non-negative Matrix Factorization (NMF) [1] is a relatively recent technique, which is used to decompose a data matrix $\boldsymbol{Y}$ into two factors $\boldsymbol{H}$ and $\boldsymbol{U}$ with non-negative entries. Owing to its simplicity and intuitive decomposition, it has attracted a lot of attention in many scientific and engineering areas in recent years. In the area of music signal processing, one successful approach involves applying NMF to a magnitude spectrogram (time-frequency representation) interpreted as a non-negative matrix [2], where the spectrogram $\boldsymbol{Y}$ is factorized into the product of a basis matrix $\boldsymbol{H}$ consisting of spectrum atoms and an activation matrix $\boldsymbol{U}$ consisting of time-varying amplitudes associated with these atoms. An important feature of this approach is that it is capable of finding a finite set of spectrum atoms that are considered to be the dominant elements constituting the observed spectrogram, in an unsupervised manner. This ability is proven to be very powerful and has enabled the NMF approach to be applied with notable success to many tasks including monaural source separation, noise reduction, music transcription, bandwidth expansion, and missing data imputation.

NMF usually considers decompositions that are approximative in nature. That is, the decomposition is performed so that the product $\boldsymbol{X} = \boldsymbol{HU}$ should approximate the original data $\boldsymbol{Y}$ as well as possible. Although the approximation error can be minimized as desired by using a larger number of spectrum atoms, the obtained basis vectors will be less likely to represent the spectra of meaningful audio events as the number of bases becomes larger. As an extreme example, if we use the same number of bases as the number of the frequency bins, we obtain an exact reconstruction $\boldsymbol{Y} = \boldsymbol{X} = \boldsymbol{HU}$ under a trivial solution $\boldsymbol{H} = \boldsymbol{I}$ (where $\boldsymbol{I}$ is an identity matrix) and $\boldsymbol{U} = \boldsymbol{Y}$, which is no longer a meaningful decomposition. On the other hand, if we use a relatively small number of bases, the chance of obtaining a meaningful decomposition will increase, but the resulting decomposition will become less accurate. This illustrates the fact that, for classical NMF, there is a trade-off between the accuracy of the reconstruction and the meaningfulness of the decomposition. To achieve an accurate approximation and thus explain the data with the NMF model as well as possible, we must use many bases. To obtain a meaningful decomposition while using many bases, we must incorporate reasonable assumptions other than non-negativity to further constrain the model $\boldsymbol{HU}$.

Another issue concerning NMF is the local optimum problem. Typically most cost functions used in NMF are difficult to optimize analytically with respect to $\boldsymbol{H}$ and $\boldsymbol{U}$. In addition, they are not jointly convex in both of the arguments $\boldsymbol{H}$ and $\boldsymbol{U}$. Thus, many existing algorithms developed for NMF are guaranteed to converge to one of the stationary points but not necessarily to a global optimum. In particular, when applying NMF to music signals, we would want to avoid local optimum solutions that are "musically" unacceptable. A good strategy to guarantee that we obtain a musically likely solution would be to incorporate music-specific constraints into the model $\boldsymbol{HU}$ or into the optimization problem.

Using these considerations as a basis, we have been concerned with developing improved variants of NMF, tailored specifically for music signals, by utilizing reasonable assumptions that we can make about music spectrograms. This paper introduces our ongoing work along with some new ideas on NMF variants incorporating music-specific constraints.

2. NMF WITH TIME-VARYING BASIS SPECTRA

2.1. Motivation

When applying the classical NMF to music spectrograms, we may expect the spectra of a single note produced by a musical instrument to be represented using a single basis spectrum scaled by time-varying amplitudes. However, its variations in time are actually much richer. A piano note would be more accurately characterized by a succession of several basis spectra corresponding to, for example, "attack", "decay", "sustain" and "release" segments. As another example, singing voices and string instruments feature a particular musical effect, vibrato, which can be characterized by its "depth" (the range of pitch variation), and its "speed" (the rate at which the pitch varies). Learning such time-varying spectra with the classical NMF would require the use of a large number of bases, and some postprocessing to group the bases into single events. However, as indicated earlier, blindly increasing the number of bases will not necessarily give meaningful decompositions at the NMF stage. If we want to increase the number of bases while maintaining the meaningfulness of the decomposition, we will need some additional constraints and/or regularization to correspondingly reduce the degree of freedom of the model. In this section, we briefly review the model described in [3, 4], which is designed to undertake the decomposition and group the basis spectra simultaneously based on the concept of time-varying basis spectra.

2.2. Model

Let us begin by dividing the bases into groups each of which we expect to correspond to a single note produced by a particular musical instrument. The NMF model $\boldsymbol{X} = (X_{\omega,t})_{\Omega\times T}$ can then be

expressed as

$$X_{\omega,t} = \sum_{k=1}^{K} \sum_{i=1}^{I_k} H_{\omega,k}^{(i)} U_{k,t}^{(i)}, \tag{1}$$

where k is the group index, i is the basis index in each group, I_k is the number of basis spectra in group k, and ω and t are frequency and time indices, respectively. Note that so far this is the same as the classical NMF model with $\sum_k I_k$ bases. We shall now introduce the following assumptions to constrain the model:

1. For each time t, exactly one basis spectrum in each group is allowed to be activated.
2. For each group k, the order in which the basis spectra appear is governed by a Markov chain.

By incorporating assumption 1, the model can be rewritten as

$$X_{\omega,t} = \sum_{k=1}^{K} H_{\omega,k}^{(Z_{k,t})} U_{k,t}, \tag{2}$$

where $Z_{k,t} \in \{1, \ldots, I_k\}$ denotes a hidden state variable indicating which basis spectrum is supposed to be activated at time t. Notice that the superscript i is dropped from U in (2) as it is no longer necessary since we are assuming $U_{k,t}^{(i)} = 0$ for $i \neq Z_{k,t}$. From assumption 2, the path of the state variables $Z_{k,1}, \ldots, Z_{k,T}$ is governed by a state transition probability $p(Z_{k,t} = a | Z_{k,t-1} = b) = \pi_{k,a,b}$.

Now, let $\mathcal{D}(x, y)$ be a discrepancy measure between x and y such that $\mathcal{D}(x, y) \geq 0$ and $\mathcal{D}(x, y) = 0$ only if $x = y$. We can then define a goodness-of-fit measure between the observed spectrogram $\boldsymbol{Y}$ and the current NMF model $\boldsymbol{X}$

$$\mathcal{J}(\boldsymbol{Y}, \boldsymbol{X}) = \sum_{\omega,t} \mathcal{D}(Y_{\omega,t}, X_{\omega,t}). \tag{3}$$

If we define $\mathcal{D}(x, y)$ as the I-divergence [5]

$$\mathcal{D}(x, y) = y \log \frac{y}{x} - (y - x), \tag{4}$$

then minimizing $\mathcal{J}(\boldsymbol{Y}, \boldsymbol{X})$ with respect to $\boldsymbol{X}$ is known to amount to maximizing the Poisson likelihood

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{\omega,t} \mathrm{Poisson}(Y_{\omega,t}; X_{\omega,t}), \tag{5}$$

where $\mathrm{Poisson}(y; x) = x^y e^{-x}/y!$. From assumption 2, $p(\boldsymbol{Z})$ is defined by

$$p(\boldsymbol{Z}) = \prod_{k} p(Z_{k,1}|\pi_{k0}) \prod_{t=2}^{T-1} p(Z_{k,t}|Z_{k,t-1}, \boldsymbol{\pi}_k), \tag{6}$$

where $\boldsymbol{\pi}_k = \{\pi_{k,a,b}\}_{1 \leq a \leq I_k, 1 \leq b \leq I_k}$.

Readers are referred to [3, 4] for detailed derivations of the parameter inference algorithms under this setting: [3] describes an algorithm for finding the Maximum A Posteriori (MAP) estimates of $\boldsymbol{H}$, $\boldsymbol{U}$ and $\boldsymbol{Z}$. [4] describes a nonparametric Bayesian formulation of the proposed NMF variant, making it possible to infer the number of groups, K, and the number of basis spectra in each group, I_k, along with the posterior distributions of $\boldsymbol{H}$, $\boldsymbol{U}$ and $\boldsymbol{Z}$.

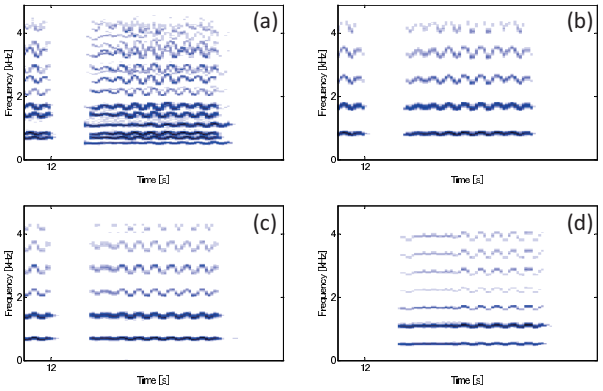

Fig. 1. Portions of the spectrograms of a mixture of 3 voices (a), estimated components corresponding to A♭ (b), F (c), and D♭ (d).

2.3. Related work

The model described above can be viewed as a factorial hidden Markov model (HMM) [6]. Recently, several authors have independently proposed modeling spectrograms using factorial HMMs [7, 8]. Let us consider a factorial HMM where the latent component generated from the k-th HMM is denoted by $C_{k,\omega,t}$, and the observed data $Y_{\omega,t}$ are assumed to be the sum of $C_{k,\omega,t}$ over k. Now, if we assume $C_{k,\omega,t} \sim \mathrm{Poisson}(C_{k,\omega,t}; H_{\omega,k}^{(Z_{k,t})} U_{k,t})$, we can confirm that the assumed factorial HMM amounts to the model described in 2.2. If we assume $C_{k,\omega,t} \sim \mathcal{N}_{\mathbb{C}}(C_{k,\omega,t}; 0, H_{\omega,k}^{(Z_{k,t})} U_{k,t})$, where $\mathcal{N}_{\mathbb{C}}(y; 0, x) \propto (1/x) e^{-|y|^2/x}$, then it amounts to the model presented in [7].

2.4. Experiment

Fig. 1 shows the decomposition achieved by the present model. Here we used a signal composed of 3 notes (D♭, F, A♭): first, each note is played alone in turn, then all two note combinations are played and finally all the notes are played simultaneously. As can be seen from Fig. 1, the present model is capable of grouping together the spectra originating from one voice even though there is a variation in pitch (vibrato).

3. NMF WITH BEAT STRUCTURE CONSTRAINT

3.1. Motivation

Music is highly structured in terms of the temporal regularity underlying the onset occurrences of notes. In general, the time between consecutive onsets corresponds to multiples and fractions of the beat period, with small deviations in timing and tempo. As we wanted each basis spectrum to correspond to a single note, the onsets of the basis spectra should have this rhythmic structure, which can be effectively used to constrain the activation matrix $\boldsymbol{U}$. In this section, we propose the introduction of a constrained variant of NMF, in which the activation matrix model is parameterized by note onsets, beat locations and tempo. To our knowledge, this is the first NMF approach that explicitly incorporates constraints derived from the rhythmic structure of music. In the following, we describe only the basic idea. For more details, please refer to [9].

3.2. Model

Let us consider a standard NMF model

$$X_{\omega,t} = \sum_{d} H_{\omega,d} U_{d,t}. \tag{7}$$

Based on the rhythmic structure of music, we make the following assumptions as regards constraining the activity function $U_{d,t}$:

1. Each activity function consists of local activity patterns, called "objects", each of which we expect to correspond to a single note activation.
2. Each object is characterized by a fast/slow rise at the onset time followed by a [1]continuous contour.
3. The onset of each object is likely to be located on the multiples or fractions of the beat period.
4. The beat period varies gradually over time.

First, from assumption 1, $U_{d,t}$ should be written as

$$U_{d,t} = \sum_{l=1}^{L_d} V_{d,l,t}, \tag{8}$$

where $V_{d,l,t}$ is the l-th object and L_d is the number of objects in the d-th activity function. To incorporate assumption 2 into $V_{d,l,t}$, we introduce a parametric model, which is expressed as the sum of Gaussians [11]

$$V_{d,l,t} = \sum_{m=1}^{M} \frac{v_{d,l} w_{d,l,m}}{\sqrt{2\pi}\phi_{d,l}} e^{-(t-(m-1)\phi_{d,l}-\tau_{d,l})^2/2\phi_{d,l}^2}, \tag{9}$$

where $v_{d,l}$, $\tau_{d,l}$ and $\phi_{d,l}$ are the total energy, onset time and parameter related to the duration of the l-th object, respectively. $w_{d,l,1}, \ldots, w_{d,l,M}$ are the weights associated with the M Gaussians that sum to unity, which determine the shape of the object. Under this constrained model $X_{\omega,t}$, we assume a Poisson likelihood as with 2.2, $p(\boldsymbol{Y}|\boldsymbol{H}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\tau}, \boldsymbol{\phi}) = \prod_{\omega,t} \mathrm{Poisson}(Y_{\omega,t}; X_{\omega,t})$. To avoid overfitting the shape of each object, we place a Dirichlet prior over $\boldsymbol{w}_{d,l} = \{w_{d,l,m}\}_{1\le m\le M}$, $p(\boldsymbol{w}) = \prod_{d,l} \mathrm{Dirichlet}(\boldsymbol{w}_{d,l}; \boldsymbol{\alpha})$, where $\mathrm{Dirichlet}(\boldsymbol{y}; \boldsymbol{x}) \propto \prod_i y_i^{x_i - 1}$ and $\boldsymbol{\alpha}$ denotes the most expected shape of the object. To promote sparsity of the activity function, we place a generalized Gaussian prior over $v_{d,l}$, $p(\boldsymbol{v}) = \prod_{d,l} \mathcal{GN}(v_{d,l}; 0, \lambda, p)$, where $\mathcal{GN}(y; 0, \lambda, p) \propto e^{-\lambda|y|^p}$ and $0 < p < 2$. Furthermore, to ensure that each basis spectrum has a harmonic structure of a particular pitch, we shall also place a Gamma prior over $\boldsymbol{H}$, $p(\boldsymbol{H}) = \prod_{\omega,d} \mathrm{Gamma}(H_{\omega,d}; \beta\bar{H}_{\omega,d}+1, \beta)$, where $\mathrm{Gamma}(x; a, b) \propto x^{a-1}e^{-bx}$.

Next, to impose Assumption 3, we first introduce a set of hyperparameters, $\boldsymbol{\psi} = \{\psi_j\}_{1\le j\le J}$, where ψ_j corresponds to the time interval between the j-th and $(j-1)$-th beat locations. With these hyperparameters, we can design a Gaussian prior distribution over the onset parameter $\tau_{d,l}$

$$p(\boldsymbol{\tau}|\boldsymbol{\psi}) = \prod_{d,l} \mathcal{N}(\tau_{d,l}; \rho_l, \nu^2), \tag{10}$$

$$\rho_l = \sum_{j=1}^{\lfloor l/I \rfloor} \psi_j + (l/I - \lfloor l/I \rfloor)\, \psi_{\lfloor l/I \rfloor + 1}, \tag{11}$$

where $\mathcal{N}(y; x, \sigma^2) \propto e^{-(y-x)^2/2\sigma^2}$, ρ_l denotes the most expected location of the onset of the l-th object, ν^2 is the variance of the Gaussian indicating how much $\tau_{d,l}$ is allowed to deviate from ρ_l, I is the number of divisions per beat, and $\lfloor x \rfloor$ denotes the largest integer not greater than x. To impose Assumption 4, it is convenient to place a Gaussian chain hyperprior over $\boldsymbol{\psi}$

$$p(\boldsymbol{\psi}) = p(\psi_1) \prod_{j=2}^{J} p(\psi_j|\psi_{j-1}), \tag{12}$$

[1]This assumption is made according to the suggestion that continuity constraints can improve the performance of NMF-based source separation [10].

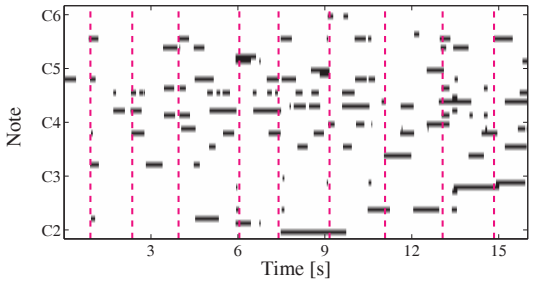

Fig. 2. Estimates of pitch, onsets/offsets, and beat locations of each note obtained with the present model applied to polyphonic data.

$$p(\psi_j|\psi_{j-1}) = \mathcal{N}(\psi_j; \psi_{j-1}, \sigma^2). \tag{13}$$

Putting altogether, the posterior density of the unknown parameters is given by

$$p(\boldsymbol{H}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\psi}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{H}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\tau}, \boldsymbol{\phi}) p(\boldsymbol{H}) p(\boldsymbol{v}) p(\boldsymbol{w}) p(\boldsymbol{\tau}|\boldsymbol{\psi}) p(\boldsymbol{\psi}). \tag{14}$$

The MAP estimates of $\boldsymbol{H}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\psi}$ can be found through an iterative algorithm consisting of parameter updates given in closed form. See [9] for its derivation.

3.3. Experiment

Fig. 2 shows an example of the application of the present model to polyphonic piano music. With the present model, we were able to estimate the pitch and onset of each note with a high detection rate, compared with the baseline NMF model [9].

4. NMF WITH TIMBRAL CLUSTERING CRITERION

4.1. Motivation

In general, each piece of music is typically played by only a handful of musical instruments or sung by one or a few singers. Thus, the constituent sounds contained in a single piece of music can probably be grouped into a reasonably small number of clusters in a certain feature space that best represents the timbral aspect of the instruments or human voice. Based on this expectation, we consider incorporating a timbral clustering criterion for $\boldsymbol{H}$ into the objective function for NMF to effectively constrain the solution space of $\boldsymbol{H}$.

4.2. Model

Let us begin by considering the standard NMF model given in (7). We are concerned with the problem of finding a decomposition such that $\boldsymbol{Y} \simeq \boldsymbol{HU}$ in which the basis spectra are forced to be clustered into some number K of clusters in a certain feature space, where we shall suppose that the value of K is given. Here we expect each group k to consist of a set of basis spectra that are timbrally consistent, meaning that all the basis spectra assigned to the same cluster are expected to be mapped onto a single point in some timbral space. Our preliminary experiments on an instrument identification task using monophonic data revealed that the mel-frequency cepstral coefficient (MFCC) is a reasonably relevant feature for robustly identifying the kinds of instruments, and so we choose here to assume the MFCC space as the feature space. The MFCC of the d-th basis spectrum is, by definition, given by

$$\mathcal{H}_{m,d} = \sum_n c_{m,n} \log \sum_\omega f_{n,\omega} H_{\omega,d}, \tag{15}$$

where $\{f_{n,\omega}\}_{1\leq\omega\leq\Omega}$ denotes the n-th triangular mel-filter in the mel-frequency filterbank and $\{c_{m,n}\}_{1\leq n\leq N}$ denotes a set consisting of the discrete cosine transform coefficients. It is worth noting that this expression reduces to the mel-spectrum when $c_{m,n}=\delta_{m,n}$ (where δ denotes Kronecker's delta), and the log-power spectrum when $f_{n,\omega}=\delta_{n,\omega}$ and $c_{m,n}=\delta_{m,n}$. For each basis spectrum d, we introduce a corresponding set of binary indicator variables $r_{d,k}\in\{0,1\}$, describing to which of the K clusters the d-th basis spectrum is assigned, so that if the d-th basis spectrum is assigned to cluster k then $r_{d,k}=1$, and $r_{d,k'}=0$ for $k'\neq k$. By using $\boldsymbol{\mu}_k$ to denote a prototype vector associated with the k-th cluster, we can define a cost function representing the sum of the distances of each feature vector to the prototype vector of the assigned cluster

$$\begin{aligned}\mathcal{R}(\boldsymbol{H},\boldsymbol{r},\boldsymbol{\mu}) &= \sum_d\sum_k r_{d,k}\left\|\boldsymbol{\mathcal{H}}_d-\boldsymbol{\mu}_k\right\|_2^2 \\ &= \sum_d\sum_k r_{d,k}\sum_m\left|\mathcal{H}_{m,d}-\mu_{m,k}\right|^2, \end{aligned} \quad (16)$$

where $\boldsymbol{r}=\{r_{d,k}\}_{1\leq d\leq D,1\leq k\leq K}$ and $\boldsymbol{\mu}=\{\boldsymbol{\mu}_k\}_{1\leq k\leq K}$. $\boldsymbol{\mathcal{H}}_d:=(\mathcal{H}_{1,d},\ldots,\mathcal{H}_{M,d})^{\mathrm{T}}$ denotes the feature vector of the d-th basis spectrum. Our goal is to find values for $\boldsymbol{H}$, $\boldsymbol{U}$, $\boldsymbol{r}$, and $\boldsymbol{\mu}$ so as to minimize

$$\mathcal{J}(\boldsymbol{Y},\boldsymbol{HU})+\lambda\mathcal{R}(\boldsymbol{H},\boldsymbol{r},\boldsymbol{\mu}), \quad (17)$$

where $\lambda>0$ is a regularization parameter. We can do this through an iterative procedure in which each iteration involves four successive steps corresponding to successive optimizations with respect to $\boldsymbol{H}$, $\boldsymbol{U}$, $\boldsymbol{r}$ and $\boldsymbol{\mu}$. Although the optimization with respect to $\boldsymbol{H}$ is mathematically intractable, we can derive a closed form update equation that guarantees a certain decrease in the objective function by using an auxiliary function approach. To do this, the first step is to define an upper bound function for the objective function (17). Owing to space limitations, we only show an upper bound function $\mathcal{R}^+$ for the regularization term $\mathcal{R}$ without proof:

$$\begin{aligned}&\mathcal{R}^+(\boldsymbol{H},\boldsymbol{r},\boldsymbol{\mu},\boldsymbol{q},\boldsymbol{\rho},\boldsymbol{\nu},\boldsymbol{\xi},\boldsymbol{\phi})\stackrel{H}{=} \quad (18)\\ &\sum_{d,k}r_{d,k}\sum_n A_{n,d}\left(\sum_\omega\frac{{\rho_{n,\omega,d}}^2}{f_{n,\omega}H_{\omega,d}}+h(\xi_{n,d})\sum_\omega f_{n,\omega}H_{\omega,d}\right)\\ &+\sum_{d,k}r_{d,k}\sum_n\mathbf{1}[B_{n,d,k}\geq 0]\frac{|B_{n,d,k}|}{\phi_{n,d}}\sum_\omega f_{n,\omega}H_{\omega,d}\\ &-\sum_{d,k}r_{d,k}\sum_n\mathbf{1}[B_{n,d,k}<0]|B_{n,d,k}|\sum_\omega\nu_{n,\omega,d}\log H_{\omega,d},\end{aligned}$$

where $\stackrel{H}{=}$ denotes equality up to a term independent of $\boldsymbol{H}$, $\mathbf{1}[\cdot]$ is the indicator function that takes the value 1 if its argument is true and 0 otherwise, $A_{n,d}=\sum_m {c_{m,n}}^2/\beta_{m,n,d}$, $B_{n,d,k}=-2\sum_m\alpha_{m,n,d}\mu_{m,k}c_{m,n}/\beta_{m,n,d}$, and $h(x)=2\log(x)/x+1/x^2$. $\alpha_{m,n,d}$ $\rho_{n,\omega,d}$, $\nu_{n,\omega,d}$, $\xi_{n,d}$ and $\phi_{n,d}$ are auxiliary variables. $\beta_{m,n,d}$ is an arbitrary parameter satisfying $\beta_{m,n,d}>0$ and $\sum_n\beta_{m,n,d}=1$. An exact bound is achieved when

$$\begin{aligned}\alpha_{m,n,d}\mu_{m,k} &= c_{m,n}\log\sum_\omega f_{n,\omega}H_{\omega,d}\\ &\quad+\beta_{m,n,d}(\mathcal{H}_{m,d}-\mu_{m,k}), \end{aligned} \quad (19)$$

$$\rho_{n,\omega,d}=\nu_{n,\omega,d}=\frac{f_{n,\omega}H_{\omega,d}}{\sum_{\omega'}f_{n,\omega'}H_{\omega',d}}, \quad (20)$$

$$\xi_{n,d}=\phi_{n,d}=\sum_\omega f_{n,\omega}H_{\omega,d}. \quad (21)$$

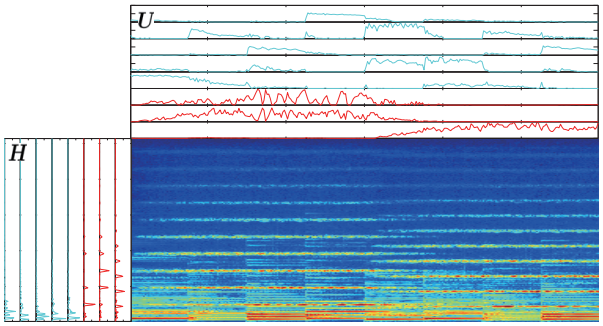

Fig. 3. An example of the decomposition obtained with the proposed regularized NMF.

4.3. Experiment

Fig. 3 shows an example of the basis spectra/activations obtained from polyphonic music played on a piano and a bass guitar (its spectrogram is shown lower right), where we assumed $K=2$. The basis spectra/activations corresponding to clusters 1 and 2 are different colors. In this example, we were able to group together the spectra originating from the same instrument automatically with the present method, even though $\boldsymbol{H}$ and $\boldsymbol{U}$ were initialized randomly.

5. SUMMARY

This paper introduced our ongoing work along with some new ideas on constrained variants of NMF that incorporate structural regularities underlying music. In the future, we plan to combine the ideas introduced in this paper to construct a unified model. While we have focused solely on the physical aspects of the regularities in music, we are also concerned with incorporating symbolic regularities, such as tonality and harmony.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for music transcription," in *Proc. WASPAA'03*, 2003, pp. 177–180.

[3] M. Nakano, J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," in *Proc. LVA/ICA'10*, 2010, pp. 149–156.

[4] M. Nakano, J. Le Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama, "Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden Markov model," in *Proc. WASPAA'11*, 2011, to appear.

[5] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.

[6] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

[7] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. WASPAA'09*, 2009, pp. 121–124.

[8] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Proc. LVA/ICA'10*, 2010, pp. 140–148.

[9] K. Ochiai, H. Kameoka, and S. Sagayama, "Explicit beat structure modeling for non-negative matrix faxtorization-based multipitch analysis," in *Proc. ICASSP'12*, 2012, to appear.

[10] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. ASLP*, vol. 15, no. 3, pp. 1066–1074, 2007.

[11] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. ASLP*, vol. 15, no. 3, pp. 982–994, 2007.