EXPLICIT BEAT STRUCTURE MODELING FOR NON-NEGATIVE MATRIX FACTORIZATION-BASED MULTIPITCH ANALYSIS

Kazuki Ochiai, Hirokazu Kameoka, Shigeki Sagayama

Graduate School of Information Science and Technology, The University of Tokyo
Hongo 7-3-1, Bunkyo, Tokyo, 113-8656, Japan

ABSTRACT

This paper proposes model-based non-negative matrix factorization (NMF) for estimating basis spectra and activations, detecting note onsets and offsets, and determining beat locations, simultaneously. Multipitch analysis is a process of detecting the pitch and onset of each note from a musical signal. Conventional NMF-based approaches often lead to unsatisfactory results very possibly due to the lack of musically meaningful constraints. As music is highly structured in terms of the temporal regularity underlying the onset occurrences of notes, we use this rhythmic structure to constrain NMF by parametrically modeling each note activation with a Gaussian mixture and derive an algorithm for iteratively updating model parameters. It is experimentally shown that the proposed model outperforms the standard NMF algorithms as regards onset detection rate.

Index Terms— Polyphonic pitch transcription, Non-negative matrix factorization, Rhythmic/Beat structure, Onset detection

1. INTRODUCTION

Music transcription is a process of obtaining a symbolic representation (such as a set of MIDI messages or a score) of an audio signal. While there are a number of viable solutions for transcribing monophonic music, polyphonic music still poses a formidable challenge.

Most existing methods for polyphonic pitch transcription rely on prior knowledge about the sources contained in the polyphonic data being analyzed. The main weakness of these kinds of methods is that they lack the capacity to adapt to signals that do not comply with the assumption they make about the sources. On the other hand, relatively recent techniques based on sparse representations use as few hypotheses as possible about the audio content to separate the notes. The goal of this kind of approach is to find a set of basis spectra such that any observed spectrum can be concisely represented as a linear combination of a small number of 'active' basis spectra. One successful approach involves applying the Non-negative Matrix Factorization (NMF) to a power spectrogram (a time-frequency representation) interpreted as a non-negative matrix [1]. In this approach, a spectrogram $\boldsymbol{Y}$ is factorized into a product of two factors with non-negative entries, one being a basis matrix $\boldsymbol{H}$, consisting of basis spectra, and the other an activation matrix $\boldsymbol{U}$, consisting of time-varying amplitudes associated with the basis spectra.

One way of obtaining a symbolic representation from a polyphonic audio signal is to apply NMF to its spectrogram and then perform onset/offset detection in the activations associated with the basis spectra [2]. This transcription method is based on the assumption that each basis spectrum obtained with NMF corresponds to a single pitch. We should thus be able to determine note onsets and offsets by simply thresholding the lines of the activation matrix $\boldsymbol{U}$. In practice, however, the lack of constraints in the NMF model often leads to unsatisfactory results. One reason is that the envelope of each basis activation typically contains many noisy peaks and dips due to the mismatch between the actual source spectra and the basis spectra. These kinds of errors occurring in the NMF phase will propagate through to the onset/offset detection phase, thus making it difficult to reach correct decisions. To obtain the basis activity information with as few errors as possible, it is necessary to incorporate an appropriate constraint into the NMF model. Many attempts have already been made to develop constrained variants of the NMF model: "Sparse NMF" promotes the sparsity of the basis activations while performing decomposition [3]. "Non-negative matrix factor deconvolution (NMFD)" [4] and "non-negative matrix factor 2D deconvolution (NMF2D)" [5] use spectro-temporal signatures rather than the basis spectra to represent spectrograms. "Smooth NMF" enforces temporal continuity on the basis activations [6]. "Harmonic NMF" imposes harmonicity constraints on the basis spectra [7]. The "Non-negative hidden Markov model" uses Markov-chained basis spectra to represent time-varying patterns in spectrograms [8–11].

In our view, transcription methods that consist of performing NMF or one of its variants mentioned above for the preprocessing, and then performing onset/offset detection for the postprocessing will at a certain point face limitations in terms of pitch transcription performance. This is because the estimation of basis activations and the detection of note onsets and offsets are each a prerequisite of the other. If we were able to obtain the correct basis activations, then it would be a relatively simple matter to detect the onsets and offsets of the underlying notes. On the other hand, if we were provided with the onsets and offsets of the notes, they could constitute very useful information for estimating the basis activations. Therefore, this leads to a "chicken and egg" situation. Furthermore, as the onsets of notes are governed by the rhythmic structure of a piece of music, the "chicken and egg" situation also applies to the detection of note onsets and the determination of beat locations. If we knew the beat locations of a piece of music, then it would be much easier to detect note onsets from the basis activations, and vice versa. Because on this, we consider it necessary to introduce a unified model for multipitch analysis, which could be used to jointly solve the problems of determining the basis spectra and activations, detecting the note onsets and offsets, and determining the beat locations.

2. PROPOSED MODEL

2.1. Non-negative Matrix Factorization model

Let us start our discussion with the standard NMF model

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

where k is the index of each basis spectrum, ω and t are frequency and time indices, respectively. The set consisting of $H_{1,k}, \ldots, H_{\Omega,k}$ represents the k-th basis spectrum and $U_{k,t}$ is

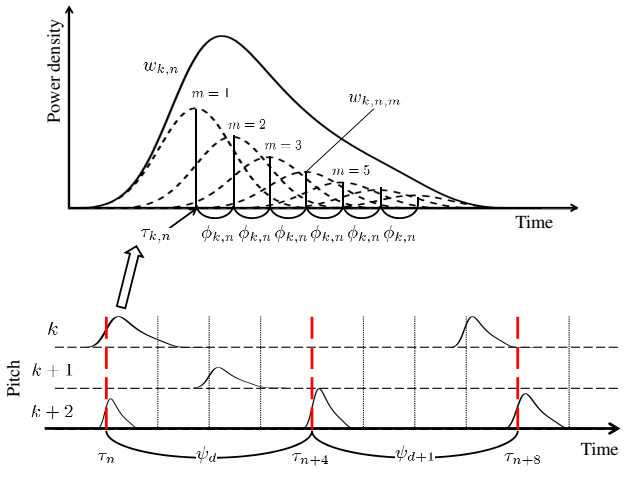

Fig. 1. An illustration of the parametric activation model, $V_{k,n,t}$, consisting of $M = 7$ Gaussians (top) and along with the beat locations (bottom). The vertical dashed lines and the dotted lines represent multiples and fractions of the beat periods.

the activity of the k-th basis spectrum at time t. Given an observed spectrogram $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega\times T}$, the goal of NMF is to find $\boldsymbol{H} = (H_{\omega,k})_{\Omega\times K} \in \mathbb{R}^{\geq 0,\Omega\times K}$ and $\boldsymbol{U} = (U_{k,t})_{K\times T} \in \mathbb{R}^{\geq 0,K\times T}$ such that $\boldsymbol{Y} \simeq \boldsymbol{X}$ where $\boldsymbol{X} = (X_{\omega,t})_{\Omega,T} = \boldsymbol{HU}$. We assume

$$\sum_{\omega} H_{\omega,k} = 1 \quad (k = 1, \ldots, K), \tag{2}$$

in order to avoid an indeterminacy in the scaling of $\boldsymbol{H}$ and $\boldsymbol{U}$.

2.2. Explicit beat structure constraint

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.

Based on the rhythmic structure of music, we make the following assumptions to constrain the activity function $U_{k,t}$:

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 continuous contour.
3. The onset of each object is likely to be located at multiples or fractions of the beat period.
4. The beat period varies gradually over time.

First, from assumption 1, $U_{k,t}$ can be expressed as

$$U_{k,t} = \sum_{n=1}^{N_k} V_{k,n,t}, \tag{3}$$

where $V_{k,n,t}$ denotes the n-th object and N_k is the number of objects in the k-th activity function. To incorporate assumption 2 into $V_{k,n,t}$, we introduce a parametric model from [12], which is expressed as a sum of Gaussians (Fig. 1):

$$V_{k,n,t} = \sum_{m=1}^{M} G_{k,n,m,t}, \tag{4}$$

$$G_{k,n,m,t} = \frac{v_{k,n} w_{k,n,m}}{\sqrt{2\pi}\phi_{k,n}} e^{-(t-(m-1)\phi_{k,n}-\tau_{k,n})^2/2\phi_{k,n}^2}, \tag{5}$$

where $v_{k,n}$ is the total energy of object n, and $\tau_{k,n}$ is the center of the first Gaussian, which can be considered an estimate of the onset time. The centers of the Gaussians are constrained to be equally spaced with the distance $\phi_{k,n}$, which is equal to the "standard deviation" of all the Gaussians. This specific constraint allows the entire object to be stretched or shrunk linearly in the time direction according to the value of $\phi_{k,n}$. Thus, $\phi_{k,n}$ can be regarded as a parameter related to the duration of a note. $w_{k,n,1}, \ldots, w_{k,n,M}$ are weights associated with the M Gaussians, which determine the shape of the object. To avoid an indeterminacy in the scaling of $v_{k,n}$ and $w_{k,n,m}$, we assume

$$\forall_{k,n} : \sum_{m=1}^{M} w_{k,n,m} = 1. \tag{6}$$

Now, recall that our goal is to achieve $\boldsymbol{Y} \simeq \boldsymbol{X}$. We thus need to define a discrepancy measure between these two quantities. Let $\mathcal{D}(y\|x)$ be a discrepancy measure between y and x such that $\mathcal{D}(y\|x) \geq 0$ and $\mathcal{D}(y\|x) = 0$ only if $y = x$. We can then define a goodness-of-fit measure between the observed spectrogram $\boldsymbol{Y}$ and the current NMF model $\boldsymbol{X}$

$$\mathcal{J}(\boldsymbol{Y}\|\boldsymbol{X}) = \sum_{\omega,t} \mathcal{D}(Y_{\omega,t}\|X_{\omega,t}). \tag{7}$$

As in [13], we choose to use the I-divergence as the distortion measure $\mathcal{D}(y\|x)$

$$\mathcal{D}(y\|x) = y \log \frac{y}{x} - (y - x). \tag{8}$$

Minimizing $\mathcal{J}(\boldsymbol{Y}\|\boldsymbol{X})$ with respect to $\boldsymbol{X}$ is then known to be equivalent to maximizing the Poisson likelihood. The likelihood of the unknown parameters, $\boldsymbol{H}$, $\boldsymbol{v}$, $\boldsymbol{w}$, $\boldsymbol{\tau}$, $\boldsymbol{\phi}$, given $\boldsymbol{Y}$ is therefore

$$p(\boldsymbol{Y}|\boldsymbol{H}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\tau}, \boldsymbol{\phi}) = \prod_{\omega,t} \mathrm{Poisson}(Y_{\omega,t}; X_{\omega,t}), \tag{9}$$

where $\mathrm{Poisson}(y; x) = x^y e^{-x}/y!$.

To avoid overfitting the shape of each object, we place a Dirichlet prior over $\boldsymbol{w}_{k,n} = \{w_{k,n,m}\}_{1\leq m\leq M}$, namely $p(\boldsymbol{w}) = \prod_{k,n} \mathrm{Dirichlet}(\boldsymbol{w}_{k,n}; \boldsymbol{\alpha})$, where $\mathrm{Dirichlet}(\boldsymbol{y}; \boldsymbol{x}) \propto \prod_i y_i^{x_i-1}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$. $\alpha_m / \sum_{m'} \alpha_{m'=1}^{M}$ is the expected value of $w_{k,n,m}$. To enforce the sparseness of the activity function, we place a generalized Gaussian prior over $v_{k,n}$, as in [14], $p(\boldsymbol{v}) = \prod_{k,n} \mathcal{GN}(v_{k,n}; 0, \lambda, p)$, where $\mathcal{GN}(y; 0, \lambda, p) \propto e^{-\lambda|y|^p}$. When $0 < p < 2$ this distribution becomes super-Gaussian and promotes the sparsity of active objects. Henceforth, we assume $0 < p \leq 1$. Furthermore, we would like to ensure that each basis spectrum has a harmonic structure of a particular pitch. To do so, we shall place a Gamma prior over $\boldsymbol{H}$, namely $p(\boldsymbol{H}) = \prod_{\omega,k} \mathrm{Gamma}(H_{\omega,k}; \gamma\bar{H}_{\omega,k} + 1, \beta)$, where $\mathrm{Gamma}(x; a, b) \propto x^{a-1}e^{-bx}$, and its mode is given by $\bar{H}_{\omega,k}$. $\bar{H}_{\omega,k}$ is thus the most likely value for $H_{\omega,k}$ and β determines the peakiness of the density around the mode.

Next, to impose Assumption 3, we first introduce a set of hyperparameters, $\boldsymbol{\psi} = \{\psi_d\}_{0\leq d\leq D}$, where ψ_d corresponds to the time interval between the $(d+1)$-th and d-th beat locations. With these hyperparameters, we can design a Gaussian prior distribution over the onset parameter $\tau_{k,n}$

$$p(\boldsymbol{\tau}|\boldsymbol{\psi}) = \prod_{k,n} \mathcal{N}(\tau_{k,n}; \rho_n, \nu^2), \tag{10}$$

$$\rho_n = \sum_{l=0}^{d-1} \psi_l + \frac{i-1}{I}\psi_d, \tag{11}$$

where $\mathcal{N}(x;\mu,\sigma^2) \propto e^{-(x-\mu)^2/2\sigma^2}$, ρ_n denotes the most expected location of the onset of the n-th object, ν^2 is the variance of the Gaussian indicating how much $\tau_{k,n}$ is allowed to deviate from ρ_n, I is the number of divisions per beat, and the indices d, i are such that $n = (d-1)I + i$. To impose Assumption 4, we place a Gaussian chain hyperprior over $\boldsymbol{\psi}$

$$p(\boldsymbol{\psi}) = p(\psi_0)\prod_{d=1}^{D} p(\psi_d|\psi_{d-1}), \tag{12}$$

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

3. PARAMETER ESTIMATION

3.1. Maximum a posteriori (MAP) estimation problem

Given an observed spectrogram $\boldsymbol{Y}$, we would like to find the estimates of $\boldsymbol{H},\boldsymbol{v},\boldsymbol{w},\boldsymbol{\tau},\boldsymbol{\phi},\boldsymbol{\psi}$ that maximize the log posterior density $p(\boldsymbol{H},\boldsymbol{v},\boldsymbol{w},\boldsymbol{\tau},\boldsymbol{\phi},\boldsymbol{\psi}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{v},\boldsymbol{w},\boldsymbol{\tau},\boldsymbol{\phi})p(\boldsymbol{H})p(\boldsymbol{v})p(\boldsymbol{w})p(\boldsymbol{\tau}|\boldsymbol{\psi})p(\boldsymbol{\psi})$. We therefore consider the problem of maximizing

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{v},\boldsymbol{w},\boldsymbol{\tau},\boldsymbol{\phi}) + \log p(\boldsymbol{H}) + \\ \log p(\boldsymbol{v}) + \log p(\boldsymbol{w}) + \log p(\boldsymbol{\tau}|\boldsymbol{\psi}) + \log p(\boldsymbol{\psi})\end{aligned}$$

subject to

$$\begin{aligned}&\forall_k : \sum_{\omega} H_{\omega,k} = 1,\ \forall_{k,n} : \sum_{m} u_{k,n,m} = 1,\ \forall_d : \psi_d \geq 0,\\ &\forall_{\omega,k} : H_{\omega,k} \geq 0,\ \forall_{k,n,m} : v_{k,n}, w_{k,n,m}, \phi_{k,n}, \tau_{k,n} \geq 0,\end{aligned}$$

where $\boldsymbol{\theta}$ denotes a set consisting of $\boldsymbol{H},\boldsymbol{v},\boldsymbol{w},\boldsymbol{\tau},\boldsymbol{\phi},\boldsymbol{\psi}$. Although it is difficult to solve this optimization problem analytically, we can develop a computationally efficient algorithm for finding a locally optimal solution based on the auxiliary function concept, similar to the one used in [12, 13].

3.2. Designing an auxiliary function

When applying an auxiliary function approach to a maximization problem, the first step is to define a lower bound function for the objective function. The difficulty with the current maximization problem lies in the terms $\log p(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{v},\boldsymbol{w},\boldsymbol{\tau},\boldsymbol{\phi})$ and $\log p(\boldsymbol{v})$. We use the following inequalities to bound these terms from below:

$$\begin{aligned}&\log p(Y_{\omega,t}|\boldsymbol{H},\boldsymbol{v},\boldsymbol{w},\boldsymbol{\tau},\boldsymbol{\phi})\\ &\geq Y_{\omega,t}\sum_{k,n,m}\gamma_{k,n,m,\omega,t}\log\frac{H_{\omega,k}G_{k,n,m,t}}{\gamma_{k,n,m,\omega,t}} - X_{\omega,t} - \log(Y_{\omega,t}!),\end{aligned}$$

$$\log p(v_{k,n}) \geq -\lambda(p\eta_{k,n}^{p-1}(v_{k,n}-\eta_{k,n})+\eta_{k,n}) + \log\frac{p\lambda^{1/p}}{2\Gamma(1/p)},$$

for $v_{k,n} > 0$, in which the exact bounds are achieved when

$$\gamma_{k,n,m,\omega,t} = \frac{H_{\omega,k}G_{k,n,m,t}}{\sum_{k',n',m'}H_{\omega,k'}G_{k',n',m',t}}, \tag{14}$$

$$\eta_{k,n} = v_{k,n}. \tag{15}$$

The auxiliary function for our objective function will thus be defined by replacing the terms $\log p(Y_{\omega,t}|\boldsymbol{H},\boldsymbol{v},\boldsymbol{w},\boldsymbol{\tau},\boldsymbol{\phi})$ and $\log p(v_{k,n})$ with the lower bounds shown above.

3.3. Parameter updates

The next step is to derive update equations for the unknown parameters using the auxiliary function. Here, we must take account of the normalization conditions for $H_{\omega,k}$ and $w_{k,n,m}$. As for $H_{\omega,k}$, below we consider a simple updating procedure, which consists of solving the unconstrained maximization of the auxiliary function and projecting its solution onto the constrained space. The update equations are as follows:

$$v_{k,n} \leftarrow A_{k,n}/(1+\lambda p\eta_{k,n}^{p-1}), \tag{16}$$

$$w_{k,n,m} \leftarrow \frac{\sum_{\omega,t}W_{k,\omega,t}G_{k,n,m,t}+\alpha_m-1}{A_{k,n}+\sum_{m'}(\alpha_{m'}-1)}, \tag{17}$$

$$\phi_{k,n} \leftarrow \frac{-b_{k,n}+\sqrt{{b_{k,n}}^2+4A_{k,n}c_{k,n}}}{2A_{k,n}}, \tag{18}$$

$$\tau_{k,n} \leftarrow \frac{\nu^2 B_{k,n}+\phi_{k,n}^2\{\Psi_d+(i-1)\psi_d/I\}}{\nu^2 A_{k,n}+\phi_{k,n}^2}, \tag{19}$$

$$\psi_d \leftarrow \frac{I\sigma^2\sum_{i,k}(i-1)(\tau_{k,n}-\Psi_d)+I^2\nu^2(\psi_{d-1}+\psi_{d+1})}{K\sigma^2\sum_i(i-1)^2+2I^2\nu^2}, \tag{20}$$

$$H_{\omega,k} \leftarrow H_{\omega,k}\frac{\sum_t Y_{\omega,t}U_{k,t}/X_{\omega,t}+\beta\bar{H}_{\omega,k}}{\sum_t U_{k,t}+\beta}, \tag{21}$$

where $W_{k,\omega,t} = Y_{\omega,t}H_{\omega,k}/X_{\omega,t}$, $\Psi_d = \sum_{l=0}^{d-1}\psi_l$ and

$$A_{k,n} = \sum_{\omega,t}W_{k,\omega,t}V_{k,n,t}, \tag{22}$$

$$B_{k,n} = \sum_{\omega,t}W_{k,\omega,t}\sum_m\{t-(m-1)\phi_{k,n}\}G_{k,n,m,t}, \tag{23}$$

$$b_{k,n} = \sum_{\omega,t}(t-\tau_{k,n})W_{k,\omega,t}\sum_m(m-1)G_{k,n,m,t}, \tag{24}$$

$$c_{k,n} = \sum_{\omega,t}(t-\tau_{k,n})^2 W_{k,\omega,t}V_{k,n,t}. \tag{25}$$

Based on the auxiliary function principle, the objective function is non-decreasing under the updates (14), (15) and (16)–(21), and so the convergence of the iterative procedure is guaranteed if each iteration involves these updates.

4. EXPERIMENTAL RESULTS

To verify the performance of our method, we evaluated the onset detection rate using a few recordings found in the RWC music database [15]. The data for each evaluation were the first 23 s, mixed down to a monaural signal and resampled to 16 kHz. The constant-Q transform was used to compute spectrograms where the time resolution, the lower bound of the frequency range, and the frequency resolution were set at 16 ms, 60 Hz and 12 cents, respectively. D and the initial value of ψ_d were set at the values obtained with [16]. K, I, N, and M were set at 74, 4, $D \times I$, and 20, respectively. ν, σ, λ were all set at 1. p, β, and α_m were set at 0.5, 1000, and $1000 \times e^{-m/3}/\sum_{m'}e^{-m'/3}$, respectively. The initial values of $v_{k,n}$, $w_{k,n,m}$, $\phi_{k,n}$, and $\tau_{k,n}$ were set at 20, $e^{-\frac{m}{3}}/\sum_{m'}e^{-\frac{m}{3}}$, 5, and $\sum_{l=0}^{d-1}\psi_l + (i-1)\psi_d/I$, respectively. The initial values of $H_{\omega,k}$ and $\bar{H}_{\omega,k}$ were set at the value obtained with the standard NMF applied to the piano excerpts from the RWC musical instrument sound database [17]. The algorithm was run for 300 iterations, after which any object whose energy $v_{n,k}/M\phi_{n,k}$ was less than a certain threshold was considered to be silent.

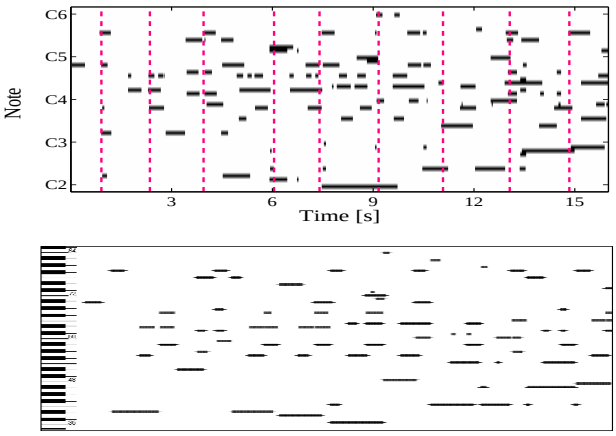

Fig. 2. An estimated piano roll (top) and the reference pitch data (bottom) for Chopin's Nocturne No. 2 in E♭-maj, Op. 9. The vertical dashed lines represent the estimated beats.

with the present method applied to Chopin's Nocturne No. 2 in E♭-maj, Op. 9 (RWC-MDB-C-2001 No. 30). We compared the proposed method with the standard NMF by employing the F-measure. We used the same reference pitch data as in [12]. The number of notes for the proposed method was simply the number of objects. As for the standard NMF, $H_{\omega,k}$ was initialized in the same manner as the proposed method, and after convergence the k-th basis spectrum was considered to be turned on at time t when $U_{k,t}$ exceeded a threshold and turned off when it was below this threshold. The threshold for each method varied from zero to the maximum value of the activations, and the maximum F-measure was selected. The results are shown in Tab. 1. It is worth noting that it is harder to obtain a high detection rate for an onset detection task than a frame-by-frame pitch detection task. To illustrate this, consider a task that involves detecting the pitch of a note with a duration of 100 frames. Suppose that method A was able to detect all the frames correctly except for one in the middle of the duration. The frame-by-frame pitch detection rate in this case will be 99%. On the other hand, the onset detection rate will be only 50% (one correct onset at the starting frame and one inserted onset in the middle of the duration). As can be seen from Tab. 1, the standard NMF model provided poor results, implying that the basis activations obtained with the standard NMF had redundant peaks and dips. On the other hand, the proposed method obtained a significantly higher detection rate especially for the piano recordings. The results for the guitar recordings were not as high as for the piano recordings because the initial basis $\boldsymbol{H}$ and the mode of the Gamma prior were set at the spectra of piano notes. If we were also provided with a set of basis spectra learned from other instruments, we would be able to obtain a higher detection rate for polyphonic music played on many kinds of instruments. Furthermore, the initial setting of the estimates of the beat locations is an important issue, that must be solved in the future, since it became clear that the estimates of the beat locations were sensitive to initialization.

5. CONCLUSION

In this paper, we proposed a new framework for the multipitch analysis of polyphonic music signals based on NMF. By incorporating the object model into the activation matrix, the onsets and the beat structure, which are very important for a proper audio-to-score music transcription, can be obtained simultaneously, together with the spectral basis matrix. We performed experiments that showed that this approach properly yields the onset times and the beat locations, and also provides higher pitch estimation accuracy than the standard NMF. Our future work will include employing time-varying spectral basis patterns to deal with the timbre change of the instruments [10], improving the estimation of the beat locations by incorporating models of musical rhythm and implementing a transcription application using this model.

Table 1. Onset detection performance of the standard NMF and proposed models. F, P and R correspond to F-measure, precision, and recall, respectively (%).

Data and Instruments	Notes	standard NMF			proposed model		
		F	P	R	F	P	R
C-2001, 30, Piano	119	18.9	13.7	30.2	**68.8**	66.4	71.4
C-2001, 35, Piano	50	10.3	5.8	44.0	**64.4**	72.5	58.0
J-2001, 1, Piano	155	15.0	8.5	59.4	**71.3**	62.7	82.6
J-2001, 2, Piano	69	11.8	6.8	43.5	**64.9**	60.8	69.6
J-2001, 6, Guitar	161	14.8	9.0	41.6	**59.3**	58.9	59.6
J-2001, 7, Guitar	99	12.3	8.2	24.2	**44.0**	48.2	40.4
J-2001, 8, Guitar	79	7.5	4.1	41.8	**30.7**	25.2	39.2
J-2001, 9, Guitar	94	11.7	7.1	33.0	**51.8**	49.5	54.3

6. REFERENCES

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

[2] N. Bertin, R. Badeau, and G. Richard, "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark," in *Proc. ICASSP*, 2011, vol. 1, pp. 65–68.

[3] S.A. Abdallah and M.D. Plumbley, "Unsupervised analysis of polyphonic music by sparse coding," *IEEE Trans. on Neural Networks*, vol. 17, no. 1, pp. 179–196, 2006.

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

[5] M. N. Schmidt and M Mørup, "Sparse non-negative matrix factor 2-D deconvolution for automatic transcription of polyphonic music," in *Proc. ICA*, 2006, pp. 700–707.

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

[7] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. ISMIR*, 2007, pp. 381–386.

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

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

[10] 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*, 2011, pp. 325–328.

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

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

[13] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2000, pp. 556–562.

[14] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. ICASSP*, 2009, pp. 3437–3440.

[15] M Goto, H Hashiguchi, T Nishimura, and R Oka, "RWC music database: Popular, classical, and jazz music database," in *Proc. ISMIR*, 2002, pp. 287–288.

[16] D. P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[17] M. Goto, "Development of the RWC music database," in *Proc. the 18th International Congress on Acoustics (ICA 2004)*, 2004, pp. I–553–556.