Multiresolutional Hierarchical Bayesian NMF for Detailed Audio Analysis of Music Performances

Takeshi Hori, Kazuyuki Nakamura and Shigeki Sagayama
Meiji University, Tokyo, Japan
E-mail: {hori,knaka,sagayama}@meiji.ac.jp

***Abstract*—In this paper, we discuss a method for a music performance detail analysis using multiresolution analysis allowing simultaneous estimation of pitch, precise onset, duration and intensity from polyphonic audio. The motivation is to obtain information that is detailed enough to develop a performance model of a human player. Characteristics of human performance can be observed as local and global tempo changes, sound intensity (volume or velocity in a MIDI), and articulations like slur and staccato. Estimation and extraction of such features from a musical audio signal in detail is useful for music information retrieval systems, automatic transcription systems, as well as automatic performance systems to train the relationship between music features and player performance. Our proposed system is based on non-negative matrix factorization (NMF) using hierarchical Bayesian inference, which is modeling harmonic and nonharmonic structures, note durations, intensities, and onset information stochastically. The estimation process comprises two steps. In the first step, variational Bayesian inference and a Gaussian mixture model is used to roughly estimate pitch onset, intensity and duration. These values are used as a prior for the second more detailed step, in which time resolution is doubled and the estimation is repeated to refine the results. The evaluation results show that the our proposed multiresolution Bayesian model can estimate more precise onset times and durations than our non-multiresolution Bayesian model.**

I. Introduction

Human music performances are influenced by a lot of factors. For example, a musician adds various expressive nuances in his performance, for example by varying the tempo or dynamics while performing. In addition to the musical intention of the composer contained in the sheet music, the interpretation of this intention by the performer, as well as the performer's own traits influence the resulting expressive performance. Additionally, in case of ensembles with multiple instruments, interactions taking into consideration each other's performance intention and instrument characteristics become important. As a result, the performances we usually listen to can vary significantly depending on the performers.

We call the mathematical consideration of various factors concerning human performance "performance engineering" in this research. Research on performance engineering has been done from various viewpoints such as automatic accompaniment, session systems, and performance expression. This paper is concerned with said performance expression, with which a human actually performs a musical score. Human performances are generally rhythmical and expressive, and not monotonous reproductions of the score as it is. The reason is that a human performer can interpret the intention of a composer from a score and vary the pace of the performance. In order to make a performance interesting and emotional, musicians usually use a multitude of dynamic and rhythmic details. For example, when a human performer plays a chord, the actual timing of the keystrokes might not be simultaneous but scattered a little. As another example, performers often emphasize the melody of a piece by slightly shifting the timing of melody notes or playing them stronger in order make it stand out. From the point of view of the listener, the deviations from the exact musical score as well as audible differences between different performers are perceived as fluctuations of onset times and durations based on the changes of global and local tempo (including articulations like slur and staccato), as well as changes of dynamics.

To properly analyze such performance properties, a high resolution of time and note intensity. We could obtain performance information from MIDI-format data [1] easily, but we would need to use a special instrument like a MIDI piano to obtain such data. Therefore, relying on MIDI data would strongly limit the amount of data we can use. However, audio recordings of human performances are available in large numbers. Thus, in order to increase the amount of data usable for performance analysis, our aim is to estimate pitches, onset times, durations and intensities of all notes from a musical audio signal. Such analysis of audio data is in the following referred to as "performance detail analysis".

In recent years, multipitch analysis has been approached with various methods [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Especially, non-negative matrix factorization (NMF) [12] was frequently used for estimation of pitches, since NMF suits music information quite well due to spectra being non-negative and the number of pitches being limited [13, 14, 15, 16, 17, 18, 19, 20]. NMF is the method of decomposing a spectrogram into a product of two lower rank matrices, one of which consists of basis vectors expressing a fundamental frequency distribution with overtones of every pitch, while the other one contains activations based on power envelopes of the corresponding basis vectors. Although incorporating acoustic models can improve estimation accuracy, the principle of uncertainty between frequency resolution and time resolution limiting the short-time Fourier transform (STFT) makes the precise estimation difficult. Especially for performance detail analysis, high time resolution is needed to extract human characteristics. As a method for tackling the problem of the uncertainty principle, the concurrent NMF (CNMF) was proposed [21]. CNMF

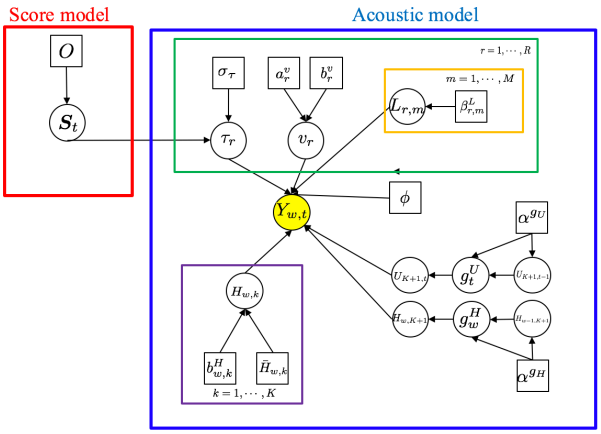

Fig. 1. Our proposed model for performance detail analysis using NMF. This model consists of an acoustic model and a score model.

is the method of using two spectrograms (high frequency resolution and high time resolution), while imposing constraints considering shared frequency and time bins of both spectrograms. Although this method estimates the pitch and precise onset successfully, detailed durations and intensities cannot be obtained.

In this paper, we propose a method for multiresolution analysis using NMF based on hierarchical Bayesian inference for detailed analysis. Our proposed method uses a model of a harmonic structure for basis matrices and approximates activation distributions of pitches with a Gaussian mixture model (GMM) as proposed in [11]. By modeling onset time information stochastically, we can utilize score information explicitly. Fig. 1 shows our basic idea. In our proposed model, the generative model of a spectrogram consists of two models; an acoustic model based on NMF and a score model that accounts for tempo differences of spectrograms. In other words, we need to estimate the parameters which maximize the likelihood of an observed spectrogram $\boldsymbol{Y}$. Furthermore, we can obtain the results of analysis with high frequency and high time resolution by utilizing onset time information estimated by initial analysis with different resolution (high frequency and low time resolution).

II. Generative Model of Musical Spectrograms

A. Problem formulation

The ultimate purpose of our system is to automatically acquire the deviation information of the human performance data relative to the music score in order to automatically turn expressionless music (e.g. digitized sheet music) into expressive performances. In other words, the aim is to estimate musical features (onset time, volume, sound length) from audio data of a human performance with high temporal resolution. However, in contrast to conventional multi-pitch analysis, the performed musical score is known to the algorithm. In performance detail analysis with high temporal resolution, the low frequency resolution caused by the uncertainty principle can be a bottleneck. To solve this problem, we propose the following 3-step analysis that utilizes score information explicitly .

1) Estimation of the onset time of each note based on score information.
2) "Coarse" estimation of onset times using the estimated onset times from step 1 as a prior.
3) "Finer" estimation of onset times using the estimated onset times from step 2.

B. Model hypothesis

In this paper, we assume the following properties of musical spectrograms, note spectrums, and note energies:

1) The frequency of every pitch is stationary, and shifts in power distribution only occur due to different durations and intensities.
2) Music consists of combinations of single tones defined by pitch and duration.
3) A spectrum of a single tone comprises a fundamental frequency distribution containing its overtones, and has a nonharmonic component.
4) The frequency distribution of the nonharmonic component is smooth and continuous.
5) The development of note magnitudes in time from onset to offset time is smooth and continuous.

C. Musical spectrogram model

A musical spectrogram is observed as a set of overlapping performed single tones. In the ideal case, if every note, for instance 88 notes in case of a piano, would correspond one-to-one to a unique spectrum, which scales linearly with magnitude, and a spectrogram of a piano performance would correspond to a sum of such note spectra, the spectrogram could be decomposed into note spectra and their magnitudes. We approximate this decomposition using NMF to factorize a spectrogram $\boldsymbol{Y} \in \mathbb{R}^{W\times T}$ into the lower rank basis matrix $\boldsymbol{H} \in \mathbb{R}^{W\times K}$ and activation matrix $\boldsymbol{U} \in \mathbb{R}^{K\times T}$ as follows:

$$\boldsymbol{Y} \approx \boldsymbol{H}\boldsymbol{U}. \tag{1}$$

To compute the low rank matrices, a distance metric is required. In our model, we use the generalized Kullback-Leibler divergence (I-divergence), which can be formulated as follows:

$$\begin{aligned} &\arg\min_{\boldsymbol{H},\boldsymbol{U}} D_{KL}[\boldsymbol{Y}||\boldsymbol{H}\boldsymbol{U}], \\ &s.t.\ \forall_k \quad H_{w,k} > 0,\ U_{k,t} > 0, \end{aligned} \tag{2}$$

which is equivalent to assuming a generative model based on the Poisson distribution:

$$Y_{w,t} \sim \mathcal{P}o(Y_{w,t}|\sum_k H_{w,k}U_{k,t}). \tag{3}$$

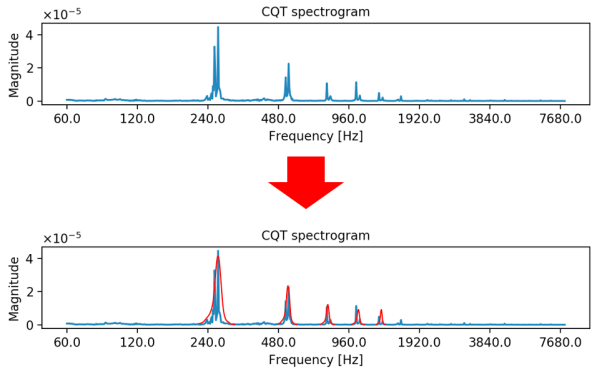

Fig. 2. An example of approximation of template basis using a GMM in the case of a recording of the note C4 played on a piano using constant-Q transform (taken from the RWC Music Database [22]).

D. Basis matrix: Harmonic component model

The basis vectors containing the respective note's harmonic components consist of a fundamental frequency and the overtone frequency distribution. As an approximation of the distribution we define the common harmonic structure pattern $h(f)$, assuming that overtone magnitudes decrease with distance from the fundamental frequency according to the following formula:

$$\frac{h(f_n)}{h(f_0)} = (n+1)^{-\alpha} \quad h(f) = 0 \text{ if } f \notin \{f_0, f_1 \ldots, f_N\} \quad (4)$$

where f_n is the frequency of the n-th overtone, and α is an attenuation coefficient. We set the parameters to $N = 8$ and $\alpha = 1.5$ for this paper. Fig. 2 illustrates the approximation of the harmonic structure. It displays the spectrum of a 'C4' note played on the piano, which can be approximated using a GMM as shown in the figure. Using relative overtone magnitudes as mixing coefficients $\gamma_{k,m}$ (m-th overtone of the k-th pitch) of the GMM, we obtain the template basis vector $\bar{H}_{w,k}$ of the corresponding note as

$$\begin{aligned} \bar{H}_{w,k} &= \sum_{m=1}^{8} \frac{\gamma_{k,m}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\log(w) - \log(mw_k))^2}{2\sigma^2}\right) \\ s.t. \quad & \sum_m \gamma_{k,m} = 1, \end{aligned} \quad (5)$$

where w_k denotes the fundamental frequency of the k-th pitch. The variance σ was set to 0.1 for this paper.

Due to the relation between the conjugate prior distribution of the Poisson distribution and the gamma distribution, the prior distribution can be expressed as

$$H_{w,k} \sim \mathcal{G}a(H_{w,k}|\alpha_{w,k}, \beta_{w,k}). \quad (6)$$

To suppress frequencies other than that of the overtones in a note's basis vector, we encourange sparsity of the basis matrix $H_{w,k}$ by using the template basis $\bar{H}_{w,k}$ as mode of the Gamma distribution,

$$H_{w,k} \sim \mathcal{G}a(H_{w,k}|(b^H_{w,k})^{-1}\bar{H}_{w,k} + 1, b^H_{w,k}). \quad (7)$$

The mode $\bar{H}_{w,k}$ can be used as the common harmonic structure pattern in (4) and (5).

E. Basis matrix: Nonharmonic component model

The nonharmonic component is mainly required to absorb the harmonic noise in a spectrogram. As the nonharmonic component has a smooth structure in the frequency dimension, the basis vector corresponding to the nonharmonic component $H_{w,K+1}$ is defined using the inverse gamma Markov chain (IGMC) [23] as

$$\begin{aligned} H_{w,K+1} &\sim \mathcal{IG}(H_{w,K+1}|\alpha^{g_H}, \frac{g^H_w}{\alpha^{g_H}}) \\ g^H_w &\sim \mathcal{IG}(g^H_w|\alpha^{g_H}, \frac{H_{w-1,K+1}}{\alpha^{g_H}}), \end{aligned} \quad (8)$$

where α^{g_H} is a hyperparameter, and g^H_w are latent auxiliary variables to ensure positive correlation between $H_{w-1,K+1}$ and $H_{w,K+1}$. Therefore, the full conditionals are

$$\begin{aligned} H_{w,K+1} &\sim \mathcal{IG}\left(H_{w,K+1}|2\alpha^{g_H}, \frac{1}{\alpha^{g_H}}\left(\frac{1}{g^H_w} + \frac{1}{g^H_{w+1}}\right)^{-1}\right) \\ g^H_w &\sim \mathcal{IG}\left(g^H_w|2\alpha^{g_H}, \frac{1}{\alpha^{g_H}}\left(\frac{1}{H_{w,K+1}} + \frac{1}{H_{w-1,K+1}}\right)^{-1}\right). \end{aligned} \quad (9)$$

F. Activation matrix: Single tone model

We assume that a musical spectrogram can be expressed as a combination of single tones. Denoting the index of an observed note as $r = \{1, \cdots, R, R+1\}$ ($r = R+1$ for the nonharmonic component) and frame numbers as $t = \{1, \cdots, T\}$, a single tone's magnitude $V_{r,t}$ can be expressed as

$$\begin{aligned} U_{k,t} &= \sum_{r=1}^{R} \delta_{\kappa_r,k} V_{r,t} \\ \delta_{\kappa_r,k} &= \begin{cases} 1 \text{ if } \kappa_r = k, \\ 0 \text{ if } \kappa_r \neq k \end{cases} \end{aligned} \quad (10)$$

where κ_r is the index of the basis vector of the corresponding note. Assuming that a single tone's magnitude $V_{r,t}$ can be described by its time-evolution $\nu_{r,t}$, which is modeled by a GMM, in combination with its energy energy v_r, $V_{r,t}$ can be rewritten as follows:

$$\begin{aligned} V_{r,t} &= v_r \nu_{r,t} \\ &= v_r \sum_{m=1}^{M} \pi_{r,m} \frac{1}{\sqrt{2\pi\phi}} \exp\left(-\frac{(t - \tau_r - (m-1)\phi)^2}{2\phi^2}\right), \\ s.t. \quad & \sum_{m=1}^{M} \pi_{r,m} = 1, \end{aligned} \quad (11)$$

where M is the number of mixtures, ϕ denotes the standard deviation, $\pi_{r,m}$ is the mixing coefficient of the GMM, and τ_r is the estimated onset time. This model fulfills the constraint of continuity of the power envelope. Furthermore, nonparametric Bayesian inference can estimate pitch duration from observed data. By applying the stick-breaking process (SBP) [24], which

is known as one of construction methods of a Dirichlet process (DP) [25], $\pi_{r,m}$ is formulated as

$$\begin{aligned} \pi_{r,m} &= L_{r,m}\prod_{l=1}^{m-1}(1-L_{r,l}) \\ L_{r,m} &\sim \mathcal{B}eta(L_{r,m}|1,\beta^L_{r,m}), \end{aligned} \tag{12}$$

where $\beta^L_{r,m}$ is a hyperparameter.

Furthermore, the activation $U_{K+1,t}$, corresponding to the nonharmonic component $H_{w,K+1}$, can also be modeled by IGMC like $H_{w,K+1}$.

$$\begin{aligned} V_{R+1,t} &= U_{K+1,t} \\ U_{K+1,t} &\sim \mathcal{IG}\left(U_{K+1,t}|2\alpha^{g_U}, \frac{1}{\alpha^{g_U}}\left(\frac{1}{g^U_t}+\frac{1}{g^U_{t+1}}\right)^{-1}\right) \\ g^U_t &\sim \mathcal{IG}\left(g^U_t|2\alpha^{g_U}, \frac{1}{\alpha^{g_U}}\left(\frac{1}{U_{K+1,t}}+\frac{1}{U_{K+1,t-1}}\right)^{-1}\right). \end{aligned} \tag{13}$$

Therefore, the activation $U_{k,t}$ can be rewritten using the single tone model and the nonharmonic component $U_{K+1,t}$ as follows.

$$\hat{U}_{k,t} = \begin{cases} \sum_{m=1}^{M} v_r \pi_{r,m} \frac{1}{\sqrt{2\pi}\phi}\exp\left(-\frac{(t-\tau_r-(m-1)\phi)^2}{2\phi^2}\right) \\ \qquad\qquad (k \neq R+1) \\ \sum_{m=1}^{M} \frac{1}{M} U_{K+1,t} \qquad (k = R+1). \end{cases} \tag{14}$$

G. Activation matrix: Single tone's energy model

As the number of notes observed in a score is limited, magnitude distribution of notes is expected to be sparse, such that the minimum number of single tones have a magnitude larger than 0. A gamma process can be used to induce sparsity for the single tone magnitude v_r,

$$v_r \sim \prod_{r=1}^{R} \mathcal{G}a(v_r|a^v_r, b^v_r), \tag{15}$$

where a^v_r and b^v_r are hyperparameters. This can also be utilized to search for performance mistakes and estimation error in the respective previous step of the multiresolution analysis.

H. Activation matrix: Onset time model

Although note onset times in human performances generally deviate from that of score-based performances (onset times according to the score) even if the performer tries to exactly follow the score, approximate time information estimated from score onset times can be used as reference information in the first step of the multiresolution analysis. We encode this information in piano-roll format using binary variables $S_{k,t}$, which are 1 if the k-th note is estimated to sound at time t. Using this approximate information in the form of r-th single tone's onset times $t_r^{S_{\kappa_r,t_0}}$ estimated using the score information for DP matching (or estimated score from previous multiresolution analysis step), the final onset time τ_r estimation is computed as follows.

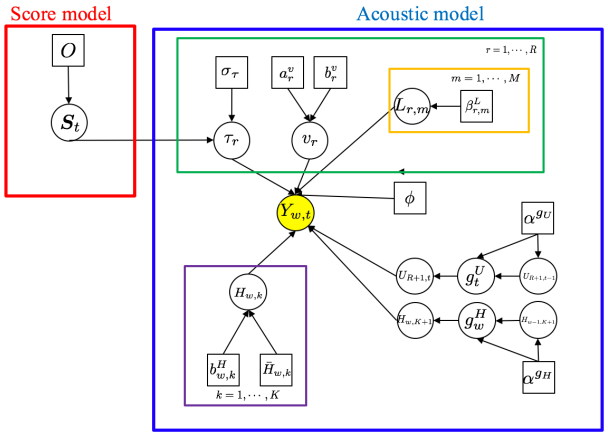

Fig. 3. Graphical model of our proposed method.

$$\tau_r \sim \mathcal{N}(\tau_r|t_r^{S_{\kappa_r,t_0}}, \sigma^2_\tau). \tag{16}$$

where the variance σ^2_τ is a hyperparameter.

I. Final formulation of the generative model

The formula of the standard NMF approximation in (3) can be rewritten using (10) and (14).

$$Y_{w,t} \sim \mathcal{P}o(Y_{w,t}|\sum_{r,m} H_{w,\kappa_r}\hat{U}_{\kappa_r,t}). \tag{17}$$

Fig. 3 illustrates the graphical model of our proposed method.

III. Approach using variational Bayesian inference

A. Variational Bayesian inference

Given the set of parameters as $\boldsymbol{\Theta}$ and the set of hyperparameters as $\boldsymbol{\Phi}$, the posterior distribution of parameters based on Section 2. and Fig. 1 can be formulated as

$$\begin{aligned} &p(\boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L}|\boldsymbol{Y}, \boldsymbol{\Phi}) \\ &\quad\propto p(\boldsymbol{Y}|\boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L}, \phi) \\ &\qquad\cdot p(\boldsymbol{H}_{1:K}|\boldsymbol{b}^H, \bar{\boldsymbol{H}})p(\boldsymbol{H}_{K+1}|\boldsymbol{g}^H, \alpha^{g_H}) \\ &\qquad\cdot p(\boldsymbol{U}_{K+1}|\boldsymbol{g}^U, \alpha^{g_U})p(\boldsymbol{v}|\boldsymbol{a}^v, \boldsymbol{b}^v) \\ &\qquad\cdot p(\boldsymbol{\tau}|\boldsymbol{S}, \sigma_\tau)p(\boldsymbol{L}|\boldsymbol{\beta}^L) \\ &\quad= p(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{\Phi}) \end{aligned} \tag{18}$$

In variational Bayesian inference, the aim is the maximization of the marginal likelihood. The lower bound of this marginal likelihood is derived using Jensen's inequality as follows.

$$\begin{aligned} \log\int_{\boldsymbol{\Theta}} p(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{\Phi})d\boldsymbol{\Theta} &\geq \int_{\boldsymbol{\Theta}} \log q(\boldsymbol{\Theta})\frac{p(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{\Phi})}{q(\boldsymbol{\Theta})}d\boldsymbol{\Theta} \\ &\equiv \mathcal{B}(q), \end{aligned} \tag{19}$$

where $q(\boldsymbol{\Theta})$ is called variational posterior distribution. In this case, the equality condition is

$$\frac{p(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{\Phi})}{q(\boldsymbol{\Theta})} = const. \tag{20}$$

So that,

$$q(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{\Phi}). \tag{21}$$

Therefore, we minimize the difference between $p(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{\Phi})$ and $q(\boldsymbol{\Theta})$, which is equivalent to maximizing the variational lower bound $\mathcal{B}(q)$ according to

$$D_{KL}[q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{\Phi})] = -\mathcal{B}(q) + \log p(\boldsymbol{Y}|\boldsymbol{\Phi}). \tag{22}$$

In variational Bayesian inference, the following mean field approximation is then used to replace the posterior distribution in order to facilitate the following derivations.

$$q(\boldsymbol{\Theta}) = \prod_i q(\theta_i) \tag{23}$$

As a result, the variational lower bound can be formulated as follows.

$$\begin{aligned} \mathcal{B}(q) &= \mathbb{E}_{q(\boldsymbol{\Theta})}[\log p(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{\Phi})] - \mathbb{E}_{q(\boldsymbol{\Theta})}[\log q(\boldsymbol{\Theta})] \\ &\propto \int_{\theta_j} q(\theta_j) \left(\mathbb{E}_{q(\boldsymbol{\Theta}_{\setminus j})}[\log p(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{\Phi})] - \log q(\theta_j) \right) d\theta_j, \end{aligned} \tag{24}$$

where $\boldsymbol{\Theta}_{\setminus j}$ denotes the set of parameters excluding the j-th parameter θ_j. (24) can be regarded as negative KL-divergence between $q(\theta_j)$ and $\mathbb{E}_{q(\boldsymbol{\Theta}_{\setminus j})}[\log p(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{\Phi})]$.

Consequently, the update equation can be formulated as follows.

$$q(\theta_j) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\Theta}_{\setminus j})}[\log p(\boldsymbol{Y}, \boldsymbol{\Theta}|\boldsymbol{\Phi})] \right). \tag{25}$$

B. The lower bound of the Poisson distribution

The expectation value of the logarithm likelihood of the Poisson distribution $\log \mathcal{P}o(Y_{w,t}|\sum_{r,m} H_{w,\kappa_r}\hat{U}_{\kappa_r,t})$ cannot be derived analytically. Jensen's inequality can be utilized for this expectation value similarly to (19).

$$\begin{aligned} &\mathbb{E}_q\left[\log \mathcal{P}o(Y_{w,t}|\sum_{r,m} H_{w,\kappa_r}\hat{U}_{\kappa_r,t})\right] \\ &= \mathbb{E}_q\left[Y_{w,t}\log\sum_{r,m} H_{w,\kappa_r}\hat{U}_{\kappa_r,t} - \sum_{r,m} H_{w,\kappa_r}\hat{U}_{\kappa_r,t}\right] \\ &\geq Y_{w,t}\sum_{r,m} C_{r,m,w,t}\mathbb{E}_q\left[\log\frac{H_{w,\kappa_r}\hat{U}_{\kappa_r,t}}{C_{r,m,w,t}}\right] \\ &\quad - \sum_{r,m}\mathbb{E}_q\left[H_{w,\kappa_r}\hat{U}_{\kappa_r,t}\right], \\ &s.t \quad \sum_{r,m} C_{r,m,w,t} = 1. \end{aligned} \tag{26}$$

The auxiliary variables $C_{r,m,w,t}$ are then derived by

$$C_{r,m,w,t} \propto \exp(\mathbb{E}_q[\log H_{w,\kappa_r}\hat{U}_{\kappa_r,t}]). \tag{27}$$

Simultaneously, the logarithm likelihood of the Poisson distribution can be used as

$$\begin{aligned} &\mathbb{E}_q\left[\log \mathcal{P}o(Y_{w,t}|\sum_{r,m} H_{w,\kappa_r}\hat{U}_{\kappa_r,t})\right] \\ &\propto \sum_{r,m}\Big(Y_{w,t}C_{r,m,w,t}\mathbb{E}_q\left[\log H_{w,\kappa_r}\hat{U}_{\kappa_r,t}\right] \\ &\quad -\mathbb{E}_q\left[\log H_{w,\kappa_r}\hat{U}_{\kappa_r,t}\right]\Big). \end{aligned} \tag{28}$$

This formula can maintain the conjugate structure with the corresponding prior distribution.

C. Variational posterior distribution

The variational posterior distribution can be expressed as follows according to the conjugate prior distributions in this model.

$$\begin{aligned} q(\boldsymbol{H}_{1:K}) &= \prod_{w,k}\mathcal{G}a(H_{w,k}|\hat{a}^H_{w,k}, \hat{b}^H_{w,k}) \\ q(\boldsymbol{H}_{K+1}) &= \prod_{w} GIG(H_{w,K+1}|\hat{a}^N_w, \hat{b}^N_w, \hat{p}^N_w) \\ q(\boldsymbol{U}_{K+1}) &= \prod_{t} GIG(U_{K+1,t}|\hat{a}^N_t, \hat{b}^N_t, \hat{p}^N_t) \\ q(\boldsymbol{v}) &= \prod_{r}\mathcal{G}a(v_r|\hat{a}^v_r, \hat{b}^v_r) \\ q(\boldsymbol{\tau}) &= \prod_{r}\mathcal{N}(\tau_r|\hat{t}^{S_{\kappa_r},t_0}_r, \hat{\sigma}^2_\tau) \\ q(\boldsymbol{L}) &= \prod_{r,m}\mathcal{B}eta(\hat{\alpha}_{r,m}, \hat{\beta}_{r,m}), \end{aligned} \tag{29}$$

where $GIG(\cdot)$ denotes the generalized inverse Gaussian distribution. In the following, the expectation $\mathbb{E}_q[\cdot]$ is replaced with $\langle\cdot\rangle$. The expectation values based on above distributions, which are used in this paper, are derived as follows:

$$\begin{aligned} \langle H_{w,k}\rangle &= \hat{a}^H_{w,k}\hat{b}^H_{w,k} \\ \langle \log H_{w,k}\rangle &= \psi(\hat{a}^H_{w,k}) + \log\hat{b}^H_{w,k} \\ \langle H_{w,K+1}\rangle &= \frac{\sqrt{\hat{b}^N_w}K_{\hat{p}^N_w+1}(\sqrt{\hat{a}^N_w\hat{b}^N_w})}{\sqrt{\hat{a}^N_w}K_{\hat{p}^N_w}(\sqrt{\hat{a}^N_w\hat{b}^N_w})} \\ \langle \log H_{w,K+1}\rangle &= \log\frac{\sqrt{\hat{b}^N_w}}{\sqrt{\hat{a}^N_w}} + \frac{\partial}{\partial\hat{p}^N_w}\log K_{\hat{p}^N_w}(\sqrt{\hat{a}^N_w\hat{b}^N_w}) \\ \langle U_{K+1,t}\rangle &= \frac{\sqrt{\hat{b}^N_t}K_{\hat{p}^N_t+1}(\sqrt{\hat{a}^N_t\hat{b}^N_t})}{\sqrt{\hat{a}^N_t}K_{\hat{p}^N_t}(\sqrt{\hat{a}^N_t\hat{b}^N_t})} \\ \langle \log U_{K+1,t}\rangle &= \log\frac{\sqrt{\hat{b}^N_t}}{\sqrt{\hat{a}^N_t}} + \frac{\partial}{\partial\hat{p}^N_t}\log K_{\hat{p}^N_t}(\sqrt{\hat{a}^N_t\hat{b}^N_t}) \\ \langle v_r\rangle &= \hat{a}^v_r\hat{b}^v_r \\ \langle \log v_r\rangle &= \psi(\hat{a}^v_r) + \log\hat{b}^v_r \\ \langle \tau_r\rangle &= \hat{t}^{S_{\kappa_r},t_0}_r \\ \langle \tau^2_r\rangle &= (\hat{t}^{S_{\kappa_r},t_0}_r)^2 + \hat{\sigma}^2_\tau \end{aligned}$$

$$\begin{aligned}\langle \log L_{r,m} \rangle &= \psi(\hat{\alpha}_{r,m}) - \psi(\hat{\alpha}_{r,m} + \hat{\beta}_{r,m}) \\ \langle \log(1 - L_{r,m}) \rangle &= \psi(\hat{\beta}_{r,m}) - \psi(\hat{\alpha}_{r,m} + \hat{\beta}_{r,m}),\end{aligned} \tag{30}$$

where K_p is a modified Bessel function of the second kind, and $\psi(\cdot)$ denotes a digamma function.

D. Derivation of $C_{r,m,w,t}$

The auxiliary variables $C_{r,m,w,t}$ are given in (27). In the case of $r \in R$,

$$\begin{aligned}&\log C_{r,m,w,t} \\ &\propto \left\langle \log H_{w,\kappa_r} \hat{U}_{\kappa_r,t} \right\rangle \\ &= \delta_{\kappa_r,k} \langle \log H_{w,k} \rangle + \langle \log v_r \rangle \\ &\quad + \langle \log L_{r,m} \rangle + \sum_{l=1}^{l-1} (\log \langle 1 - L_{r,m} \rangle) \\ &\quad - \frac{1}{2} \log 2\pi\phi^2 - \frac{1}{2\phi^2} \langle \tau_r^2 \rangle \\ &\quad + \frac{1}{\phi^2} (t - (m-1)\phi) \langle \tau_r \rangle \\ &\quad - \frac{1}{2\phi^2} (t - (m-1)\phi)^2.\end{aligned} \tag{31}$$

On the other hand, in the case of $r = R + 1$,

$$\begin{aligned}\log C_{R+1,m,w,t} &\propto \left\langle \log H_{w,\kappa_r} \hat{U}_{\kappa_r,t} \right\rangle \\ &= \left\langle \log \frac{1}{M} H_{w,K+1} U_{K+1,t} \right\rangle \\ &= \langle \log H_{w,K+1} \rangle + \langle \log U_{K+1,t} \rangle - \log M\end{aligned} \tag{32}$$

E. Derivation of $H_{w,k}$

The variational posterior distribution of the harmonic components of the basis matrix $H_{w,k}$ can be derived from the other parameters as follows.

$$\begin{aligned}&\log q(H_{w,k}) \\ &\propto \langle \log \left(p(\boldsymbol{Y} | \boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L}, \phi) \right. \\ &\quad \left. \cdot\, p(\boldsymbol{H}_{1:K} | \boldsymbol{b}^H, \bar{\boldsymbol{H}}) \right) \rangle \\ &= \left(\sum_{r,m,t} \delta_{\kappa_r,k} Y_{w,t} C_{r,m,w,t} + (b^h_{w,k})^{-1} \bar{H}_{w,k} + 1 - 1 \right) \\ &\quad \cdot \log H_{w,k} - \left(\sum_{r,m,t} \delta_{\kappa_r,k} \langle X_{r,m,t} \rangle + (b^H_{w,k})^{-1} \right) H_{w,k},\end{aligned} \tag{33}$$

where $\langle X_{r,m,t} \rangle$ is

$$\langle X_{r,m,t} \rangle = \left\langle \frac{v_r \pi_{r,m}}{\sqrt{2\pi}\phi} \exp\left(-\frac{(t - \tau_r - (m-1)\phi)^2}{2\phi^2} \right) \right\rangle \tag{34}$$

Since the integral of the Gaussian mixture distribution becomes 1 if integrating over time T and all mixture modes M, the following holds.

$$\sum_{r,m,t} \langle X_{r,m,t} \rangle = \sum_r \langle v_r \rangle. \tag{35}$$

Therefore, the parameters of the variational posterior distribution of the basis matrix of the harmonic components $q(\boldsymbol{H}_{1:K}) = \prod_{w,k} \mathcal{G}a(H_{w,k} | \hat{a}^H_{w,k}, \hat{b}^H_{w,k})$ are derived as

$$\begin{aligned}\hat{a}^H_{w,k} &= \sum_{r,m,t} \delta_{\kappa_r,k} Y_{w,t} C_{r,m,w,t} + (b^h_{w,k})^{-1} \bar{H}_{w,k} + 1 \\ \hat{b}^H_{w,k} &= \left(\sum_r \delta_{\kappa_r,k} \langle v_r \rangle + (b^H_{w,k})^{-1} \right)^{-1}.\end{aligned} \tag{36}$$

The template basis matrix $\bar{\mathbf{H}}$ is defined in (5).

F. Derivation of $H_{w,K+1}$

The variational posterior distribution of the nonharmonic component in the basis matrix H^N_w can be derived from the other parameters as follows.

$$\begin{aligned}&\log q(H_{w,K+1}) \\ &\propto \langle \log \left(p(\boldsymbol{Y} | \boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L}, \phi) \right. \\ &\quad \left. \cdot\, p(\boldsymbol{H}_{K+1} | \boldsymbol{g}^H, \alpha^{g_H}) \right) \rangle \\ &= \left(\sum_{m,t} Y_{w,t} C_{R+1,m,w,t} - 2\alpha^{g_H} - 1 \right) \log H_{w,K+1} \\ &\quad - \frac{1}{2} \left(2 \left(\sum_t \langle U_{K+1,t} \rangle \right) H_{w,K+1} \right. \\ &\qquad \left. + 2 \left\langle \left(\frac{\hat{g}^H_w}{\alpha^{g_H}} \right)^{-1} \right\rangle \frac{1}{H_{w,K+1}} \right),\end{aligned} \tag{37}$$

where

$$\left\langle \frac{1}{\hat{g}^H_w} \right\rangle = \left\langle \frac{1}{g^H_{w+1}} \right\rangle + \left\langle \frac{1}{g^H_w} \right\rangle. \tag{38}$$

Therefore, the parameters of the variational posterior distribution of the basis matrices of the nonharmonic components $q(\boldsymbol{H}_{K+1}) = \prod_w GIG(H_{w,K+1} | \hat{a}^N_w, \hat{b}^N_w, \hat{p}^N_w)$ are derived as

$$\begin{aligned}\hat{a}^N_w &= 2 \sum_t \langle U_{K+1,t} \rangle \\ \hat{b}^N_w &= 2\alpha^{g_H} \left(\left\langle \frac{1}{g^H_{w+1}} \right\rangle + \left\langle \frac{1}{g^H_w} \right\rangle \right) \\ \hat{p}^N_w &= \sum_{m,t} Y_{w,t} C_{R+1,m,w,t} - 2\alpha^{g_H}.\end{aligned} \tag{39}$$

In addition, the expectation $\langle x^{-1} \rangle$ of the inverse gamma distribution $\mathcal{IG}(x | \alpha, \beta)$ equals $\alpha\beta$.

G. Derivation of $U_{K+1,t}$

The variational posterior distribution of the nonharmonic component of an activation $U_{K+1,t}$ can be derived like

$H_{w,K+1}$.

$$\begin{aligned}
&\log q(U_{K+1,t}) \\
&\quad \propto \langle \log \left(p(\boldsymbol{Y}|\boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L}, \phi) \right. \\
&\qquad \left. \cdot\, p(\boldsymbol{U}_{K+1}|\boldsymbol{G}^U, \alpha^{g_U})\right)\rangle \\
&\quad = \left(\sum_{m,w} Y_{w,t} C_{R+1,m,w,t} - 2\alpha^{g_U} - 1\right) \log U_{K+1,t} \\
&\qquad - \frac{1}{2}\left(2\left(\sum_w \langle H_{w,K+1}\rangle\right) U_{K+1,t} \right. \\
&\qquad\qquad \left. + 2\left\langle \left(\frac{\hat{g}_t^U}{\alpha^{g_U}}\right)^{-1} \right\rangle \frac{1}{U_{K+1,t}}\right),
\end{aligned} \tag{40}$$

where

$$\left\langle \frac{1}{\hat{g}_t^U} \right\rangle = \left\langle \frac{1}{g_{t+1}^U} \right\rangle + \left\langle \frac{1}{g_t^U} \right\rangle. \tag{41}$$

Therefore, the parameters of the variational posterior distribution of an activation of nonharmonic components $q(\boldsymbol{U}_{K+1}) = \prod_t GIG(U_{K+1,t}|\hat{a}_t^N, \hat{b}_t^N, \hat{p}_t^N)$ are derived as

$$\begin{aligned}
\hat{a}_t^N &= 2\sum_w \langle H_{w,K+1}\rangle \\
\hat{b}_t^N &= 2\alpha^{g_U}\left(\left\langle \frac{1}{g_{t+1}^U}\right\rangle + \left\langle \frac{1}{g_t^U}\right\rangle\right) \\
\hat{p}_t^N &= \sum_{m,w} Y_{w,t} C_{R+1,m,w,t} - 2\alpha^{g_U}.
\end{aligned} \tag{42}$$

H. Derivation of v_r

Using (35), the variational posterior distribution of the energy of a single tone v_r is derived as

$$\begin{aligned}
&\log q(v_r) \\
&\quad \propto \langle \log \left(p(\boldsymbol{Y}|\boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L}, \phi) \right. \\
&\qquad \left. \cdot\, p(\boldsymbol{v}|\boldsymbol{a}^v, \boldsymbol{b}^v)\right)\rangle \\
&\quad = \left(\sum_{m,w,t} Y_{w,t} C_{r,m,w,t} + a_r^v - 1\right) \log v_r \\
&\qquad - \left(\sum_w \delta_{\kappa_r,k} \langle H_{w,k}\rangle + (b_r^v)^{-1}\right) v_r,
\end{aligned} \tag{43}$$

Therefore, the parameters of the variational posterior distribution of the energy of a single tone $q(\boldsymbol{v}) = \prod_r \mathcal{G}a(v_r|\hat{a}_r^v, \hat{b}_r^v)$ are derived as

$$\begin{aligned}
\hat{a}_r^v &= \sum_{m,w,t} Y_{w,t} C_{r,m,w,t} + a_r^v \\
\hat{b}_r^v &= \left(\sum_w \delta_{\kappa_r,k} \langle H_{w,k}\rangle + (b_r^v)^{-1}\right)^{-1}.
\end{aligned} \tag{44}$$

I. Derivation of τ_r

Using (35), the variational posterior distribution of the estimated onset time of the r-th note τ_r is derived as

$$\begin{aligned}
&\log q(\tau_r) \\
&\propto \langle \log \left(p(\boldsymbol{Y}|\boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L}, \phi) \right. \\
&\qquad \left. \cdot\, p(\boldsymbol{\tau}|\boldsymbol{S}, \sigma_\tau)\right)\rangle \\
&= -\frac{1}{2}\left(\frac{\sum_{m,w,t} Y_{w,t} C_{r,m,w,t}}{\phi^2} + \frac{1}{\sigma_\tau^2}\right)\tau_r^2 \\
&\quad + \left(-\frac{\sum_{m,w,t} Y_{w,t} C_{r,m,w,t}((m-1)\phi - t)}{\phi^2} + \frac{t_r^{S_{\kappa_r},t_0}}{\sigma_\tau^2}\right)\tau_r.
\end{aligned} \tag{45}$$

Using the quadratic formula

$$-\frac{1}{2}AX^2 + BX = -\frac{A}{2}(X - A^{-1}B)^2 + \frac{1}{2}A^{-1}B^2, \tag{46}$$

Since the mean is equivalent to $A^{-1}B$ and the variance is A^{-1}, the parameters of the variational posterior distribution of onset times $q(\boldsymbol{\tau}) = \prod_r \mathcal{N}(\tau_r|\hat{t}_r^{S_{\kappa_r},t_0}, \hat{\sigma}_\tau^2)$ are derived as

$$\begin{aligned}
\hat{t}_r^{S_{\kappa_r},t_0} &= A^{-1}B \\
&= \frac{\phi^2 t_r^{S_{\kappa_r},t_0}}{\sigma_\tau^2 \sum_{m,w,t} Y_{w,t} C_{r,m,w,t} + \phi^2} \\
&\quad - \frac{\sigma_\tau^2 \sum_{m,w,t} Y_{w,t} C_{r,m,w,t}\ ((m-1)\phi - t)}{\sigma_\tau^2 \sum_{m,w,t} Y_{w,t} C_{r,m,w,t} + \phi^2} \\
\hat{\sigma}_\tau^2 &= A^{-1} \\
&= \frac{\phi^2 \sigma_\tau^2}{\sigma_\tau^2 \sum_{m,w,t} Y_{w,t} C_{r,m,w,t} + \phi^2}
\end{aligned} \tag{47}$$

J. Derivation of $L_{r,m}$

The variational posterior distribution of the SBP parameter $L_{r,m}$ is derived as

$$\begin{aligned}
&\log q(L_{r,m}) \\
&\quad \propto \langle \log \left(p(\boldsymbol{Y}|\boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L}, \phi) \right. \\
&\qquad \left. \cdot\, p(\boldsymbol{L}|\boldsymbol{\beta}^L)\right)\rangle \\
&\quad = \left(\sum_{w,t} Y_{w,t} C_{r,m,w,t} + 1 - 1\right) \log L_{r,m} \\
&\qquad + \left(\sum_{w,t} \sum_{l=m+1}^{M} Y_{w,t} C_{r,l,w,t} + \beta_{r,m} - 1\right) \log(1 - L_{r,m}).
\end{aligned} \tag{48}$$

Therefore, the parameters of the variational posterior distribution of SBP parameters $q(\boldsymbol{L}) = \prod_{r,m} \mathcal{B}eta(\hat{\alpha}_{r,m}, \hat{\beta}_{r,m})$ are derived as

$$\begin{aligned}
\hat{\alpha}_{r,m} &= \sum_{w,t} Y_{w,t} C_{r,m,w,t} + 1 \\
\hat{\beta}_{r,m} &= \sum_{w,t} \sum_{l=m+1}^{M} Y_{w,t} C_{r,l,w,t} + \beta_{r,m}.
\end{aligned} \tag{49}$$

K. Variational lower bound

The variational lower bound $\mathcal{B}(q)$ (see (19)) can be rewritten as

$$\begin{aligned}\mathcal{B}(q) =& \mathbb{E}_q\left[p(\boldsymbol{Y}, \boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L}|\boldsymbol{\Phi})\right] \\ & - \mathbb{E}_q\left[q(\boldsymbol{H}_{1:K}, \boldsymbol{H}_{K+1}, \boldsymbol{U}_{K+1}, \boldsymbol{v}, \boldsymbol{\tau}, \boldsymbol{L})\right].\end{aligned} \tag{50}$$

Using the parameters of the variational posterior distribution derived in (31)-(49), this lower bound can be calculated.

IV. Experimental Evaluation

A. Experimental setup

We conducted the following experiment to evaluate the estimation accuracy of musical features (onset times and intensities) using our proposed model. However, since it is difficult to correctly estimate note duration if the sustain pedal is used, note duration was excluded from the evaluation in this study. Firstly, we computed the accuracy of the single resolution architecture with a STFT time frame length of 32ms and half-overlap time shift. The accuracy of our proposed multiresolution analysis method was computed as well and compared with the above accuracy. The analysis steps of the multiresolution architecture were as follows (see Fig. 4).

1) STFT of the audio signal of both score-based and expressive performances using a time frame length of 64ms.
2) Fast dynamic time warping (fast DTW) [26] between them to align the two performances and to estimate the onset times.
3) Application of the Hierarchical Bayesian NMF to estimate every onset time.
4) STFT of the expressive performance audio signal using a frame length of 32ms.
5) Using the obtained onset times as prior for the second NMF computation.

We used three pieces ((i) Chopin, Ballade Op.52, No.4, (ii) Chopin, Prelude Op. 28, No. 24, (iii) Chopin, Etude Op. 10, No. 1) from the International Piano-e-Competition data [27] as expressive performances, and data from the Classical Archives [28] as score-based performances for this experiment. The MIDI data was converted to monaural WAVE format using Cubase (Helion SE Piano), and the first 15 seconds of the three pieces were used for this paper.

Table I shows the parameters that we used in the experiment.

B. Experimental results

The experimental results of our proposed system are shown in Table II. 'Single' of Table II indicates the results of non-multiresolution analysis (using only score-based performance data for onset estimation). 'Multi' indicates the results of multiresolution analysis (first step using score-based performance data, second step using estimated onset times from previous step). The mean and the variance of the estimation accuracy were calculated based on the difference between correct onset times and estimated onset times. Furthermore, the intensities were evaluated based on the correlation between

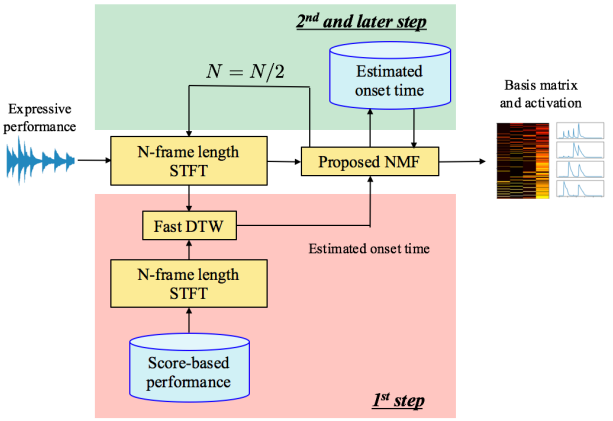

Fig. 4. Processing steps of the musical spectrogram analysis system based on multiresolutional hierarchical Bayesian NMF. In the 1st step, the system is using the score-based performance data, aligning it with the expressive performance for a first estimation of onset times. In the 2nd and later steps, the system refines the estimates from the respective previous step by using STFT with increased time resolution.

TABLE I
Parameters used for the multiresolutional hierarchical Bayesian NMF

Sampling frequency	16 kHz
Overtone magnitude attenuation factor	$\alpha = 1.5$
Number of harmonics	$N = 8$
Harmonic basis	$b^H_{w,k} = \frac{1}{3\times(\max \sum_t Y_{w,t})}$
Nonharmonic basis	$\alpha^{gH} = 0.0002$
Nonharmonic activation	$\alpha^{gU} = 0.01$
Single tone energy	$a^v_r = 1.0$, $b^v_r = 1.0$
Number of mixtures of the GMM	$M = 30$
Standard deviation of the GMM	$\phi = 1.0$
Onset variance	$\sigma^2_\tau = 2.0$
SBP	$\beta_{r,m} = 10 \times \frac{\exp(-\frac{m}{8})}{\sum_l \exp(-\frac{m}{8})}$

the normalized MIDI velocities of the MIDI data used to generate the audio data and the normalized MIDI velocities estimated from this data. Although both analysis methods (non-multiresolution and multiresolution) could estimate the onset times, the multiresolution analysis displays further increased accuracy.

Additionally, Fig. 5, 6, and7 show the piano roll data estimated from the activations, obtained using a threshold which was found by trial-and-error. For note duration estimation, we utilized the posterior parameters of the stick-breaking process $\boldsymbol{L}$ formulated in (12). We then calculated the r-th single tone's duration d_r in number of frames as

$$\begin{aligned}&\arg\max_{d_r} \sum_{m=1}^{d_r} \pi_{r,m} \\ &s.t. \quad \sum_{m=1}^{d_r} \pi_{r,m} < 0.97\end{aligned} \tag{51}$$

Piano keystrokes that look continuous in the figures can be separated using the corresponding estimated onset information explicitly.

TABLE II
MEAN AND VARIANCE OF ESTIMATED ERROR OF THE ACTIVATION

Music piece	Onset time error	Intensity correlation
(i) Op. 52, No. 4 (Single)	$\mu = -2.82$, $\sigma^2 = 22.6$	$r = 0.65$
(i) Op. 52, No. 4 (Multi)	$\mu = 0.35$, $\sigma^2 = 18.4$	$r = 0.67$
(ii) Op. 28, No. 24 (Single)	$\mu = 0.82$, $\sigma^2 = 2.13$	$r = 0.65$
(ii) Op. 28, No. 24 (Multi)	$\mu = 0.52$, $\sigma^2 = 1.04$	$r = 0.58$
(iii) Op. 10, No. 1 (Single)	$\mu = -1.63$, $\sigma^2 = 25.6$	$r = 0.31$
(iii) Op. 10, No. 1 (Multi)	$\mu = -1.56$, $\sigma^2 = 20.0$	$r = 0.38$

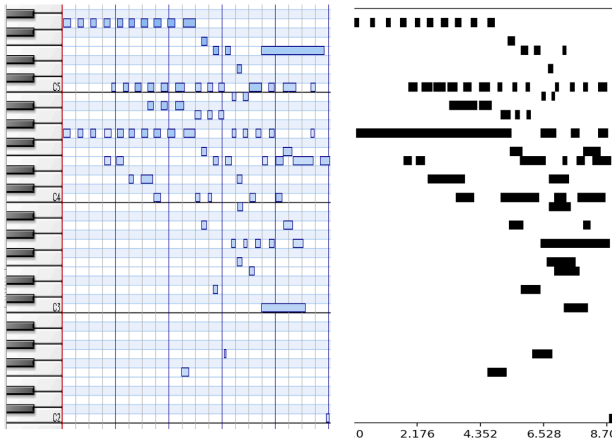

Fig. 5. The estimated piano roll from 'Chopin, Ballade Op.52, No.4'. The correct piano roll is shown on the left, and the estimated piano roll on the right. Using the corresponding estimated onset information, we can separate consecutive keystrokes of the same note even if they look continuous in the piano roll.

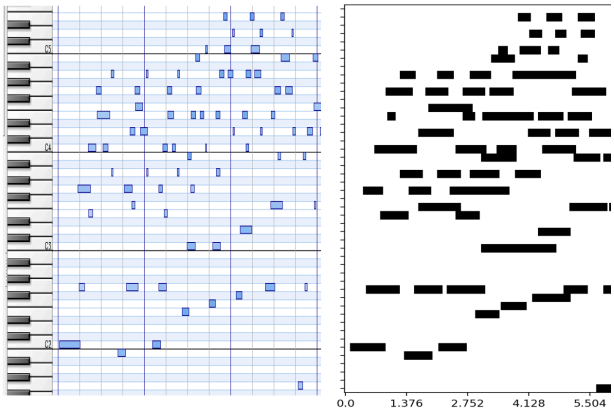

Fig. 6. The estimated piano roll from 'Chopin, Prelude Op. 28, No. 24'. The correct piano roll is shown on the left, and the estimated piano roll on the right. Using the corresponding estimated onset information, we can separate consecutive keystrokes of the same note even if they look continuous in the piano roll.

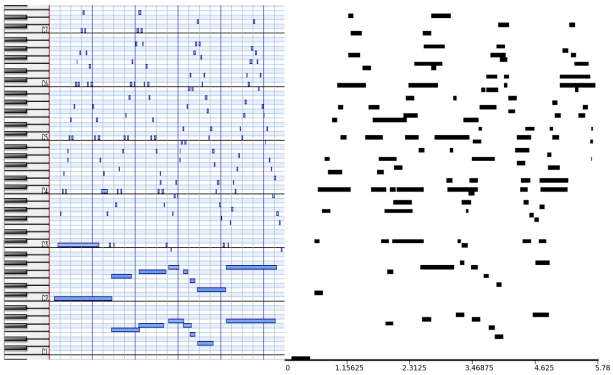

Fig. 7. The estimated piano roll from 'Chopin, Etude Op. 10, No. 1'. The correct piano roll is shown on the left, and the estimated piano roll on the right. Using the corresponding estimated onset information, we can separate consecutive keystrokes of the same note even if they look continuous in the piano roll.

Regarding the individual estimation accuracy of all onset times, although we could obtain high accuracy for several notes, there were some estimated onset times that deviated significantly from the groundtruth. In consequence, the variances of the estimation error is relatively large. The following reasons might cause this estimation error:

- Large deviations of the initial estimates obtained by time alignment with score data using DTW
- A sub-optimal choice of the variance for estimation of onset time σ_τ^2
- A sub-optimal choice of the standard deviation ϕ and the number of mixtures M of the GMM modeling the shapes of the single tones' energy envelopes.

These problems could be resolved by modeling the relationship between score-based and expressive performance spectrogram (time contraction and dilation) using not DTW but a hidden Markov model (HMM).

Consequently, the experimental results show that the multiresolution NMF analysis performed significantly better than a non-multiresolution NMF analysis for detailed analysis of musical spectrograms.

V. CONCLUSION

We proposed multiresolution NMF based on hierarchical Bayesian NMF for performance detail analysis. By applying stick-breaking processes, which are one of the construction methods of Dirichlet processes, we could obtain note durations as well as onset times and intensities of every single tone from audio data. The experimental results show that the multiresolution analysis and the hierarchical Bayesian inference is effective for performance detail analysis. In future research, applying the Beta process [20] in our system can be considered in order to obtain the piano roll (binary variables) of expressive performances, as opposed to thresholding activation values. Furthermore, individual notes in chords are usually slightly shifted in time, even if they appear simultaneously in the musical score. Therefore, considering that every note should be treated independently, we need to not only use DTW but also a factorial HMM (FHMM) that can deal with multiple independent series. Furthermore, we want to develop a human performance model, which could be used to turn expression-

less score data into expressive human-like performances.

A large amount of symbolic music data is needed for machine learning in recent years. Our proposed model can be used to obtain such data from audio recordings by estimating piano roll representations of real performances. Practically, the method could also be used without score information by using standard NMF in the first step. Our proposed architecture using multiresolution analysis can raise the estimation accuracy significantly, and we therefore think that other NMF models can also utilize this architecture in order to deal with problems like sound source separation.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 17H00749.

REFERENCES

[1] M. Hashida, T. Matsui, and H. Katayose, "A new music database describing deviation information of performance expressions." 2008.

[2] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the em algorithm for co-channel speech separation," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2. IEEE, 1993, pp. 728–731.

[3] A. Klapuri, T. Virtanen, and J.-M. Holm, "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals," in *Proc. COST-G6 Conference on Digital Audio Effects*, 2000, pp. 233–236.

[4] M. Goto, "A predominant-f/sub 0/estimation method for cd recordings: Map estimation using em algorithm for adaptive tone models," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 5. IEEE, 2001, pp. 3365–3368.

[5] S. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2. IEEE, 2002, pp. II–1769.

[6] C. Raphael, "Automatic transcription of piano music." in *ISMIR*, 2002.

[7] K. Takahashi, T. Nishimoto, and S. Sagayama, "Multi - pitch analysis using deconvolution of log - frequency spectrum," *IPSJ SIG Technical Reports*, vol. 2003, no. 127 (2003-MUS-053), pp. 61–66, 2003.

[8] M. Goto, "A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[9] H. Katmeoka, T. Nishimoto, and S. Sagayama, "Separation of harmonic structures based on tied gaussian mixture model and information criterion for concurrent sounds," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 4. IEEE, 2004, pp. iv–iv.

[10] S. Saito, H. Kameoka, T. Nishimoto, and S. Sagayama, "Specmurt analysis of multi-pitch music signals with adaptive estimation of common harmonic structure." in *ISMIR*, 2005, pp. 84–91.

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

[12] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[13] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* IEEE, 2003, pp. 177–180.

[14] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2004, pp. 494–499.

[15] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in models for acoustic processing, neural information processing systems workshop*. Citeseer, 2006.

[16] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music." in *ICML*, 2010, pp. 439–446.

[17] 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 *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 149–156.

[18] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[19] 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 *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 2011, pp. 325–328.

[20] D. Liang and M. D. Hoffman, "Beta process non-negative matrix factorization with stochastic structured mean-field variational inference," *arXiv preprint arXiv:1411.1804*, 2014.

[21] K. Ochiai, M. Nakano, N. Ono, and S. Sagayama, "Concurrent non-negative matrix factorization using multi-resolution spectrograms for multipitch analysis of music signals," *IPSJ SIG Technical Reports (MUS)*, vol. 2011, no. 5, pp. 1–6, 2011.

[22] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database : Database of copyright-cleared musical pieces and instrument sounds for research purposes," *IPSJ Journal*, vol. 45, no. 3, pp. 728–738, 2004.

[23] A. T. Cemgil and O. Dikmen, "Conjugate gamma markov random fields for modelling nonstationary sources," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 697–705.

[24] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica sinica*, pp. 639–650, 1994.

[25] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[26] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[27] International piano-e-competition. [Online]. Available: http://www.piano-e-competition.com/

[28] Classical archives. [Online]. Available: https://www.classicalarchives.com/