CONTEXT-FREE 2D TREE STRUCTURE MODEL OF MUSICAL NOTES FOR BAYESIAN MODELING OF POLYPHONIC SPECTROGRAMS

Hirokazu Kameoka[1,2], **Kazuki Ochiai**[1], **Masahiro Nakano**[2], **Masato Tsuchiya**[1], **Shigeki Sagayama**[1]
[1]Graduate School of Information Science and Technology, The University of Tokyo
Hongo 7-3-1, Bunkyo, Tokyo, 113-8656, Japan
[2]NTT Communication Science Laboratories, NTT Corporation
Morinosato Wakamiya 3-1, Atsugi, Kanagawa, 243-0198, Japan

ABSTRACT

This paper proposes a Bayesian model for automatic music transcription. Automatic music transcription involves several subproblems that are interdependent of each other: multiple fundamental frequency estimation, onset detection, and rhythm/tempo recognition. In general, simultaneous estimation is preferable when several estimation problems have chicken-and-egg relationships. This paper proposes modeling the generative process of an entire music spectrogram by combining the sub-process by which a musically natural tempo curve is generated, the sub-process by which a set of note onset positions is generated based on a 2-dimensional tree structure representation of music, and the sub-process by which a music spectrogram is generated according to the tempo curve and the note onset positions. Most conventional approaches to music transcription perform note extraction prior to structure analysis, but accurate note extraction has been a difficult task. By contrast, thanks to the combined generative model, the present method performs note extraction and structure estimation simultaneously and thus the optimal solution is obtained within a unified framework. We show some of the transcription results obtained with the present method.

1. INTRODUCTION

Music transcription is the process of automatically converting a given audio signal into a musical score. Although there are a number of viable ways of transcribing monophonic music, polyphonic music still poses a formidable challenge.

Several subproblems must be solved if we are to transcribe polyphonic music automatically, namely source separation, multiple fundamental frequency estimation (the estimation of the fundamental frequencies of concurrent musical sounds), onset detection (the detection of the position in the signal where each note begins), and rhythm recognition (the estimation of the tempo, beat locations, and the note value of each note). The difficulty is that these subproblems involve many ambiguities.

An audio signal of a musical note typically consists of many overtones, some of which usually overlap when multiple notes are played simultaneously. To detect which notes are present at a certain time instant, we need to know which musical note each frequency component belongs to. Since this information is missing for the spectrum of a mixture signal, there can be multiple interpretations of how the spectrum of each sound should appear as well as which pitches are present in the mixture. On the other hand, a music performance often involves temporal fluctuation in terms of both rhythm and tempo, which means performers do not always play notes with a perfectly timed rhythm and constant tempo. Since we cannot define a note value without having a notion for tempo and vice versa, there can be infinite interpretations regarding what the intended rhythm was and how the tempo varied if both types of information are missing.

Many methods have already been developed for polyphonic music transcription, most of which try to tackle the problem by dealing with the abovementioned subproblems separately [1]. However, the inherent difficulty of the music transcription problem lies in the chicken-and-egg interdependency between these subproblems [2]. Firstly, if the given signal is already decomposed into individual notes, it is a simple matter to detect their fundamental frequencies. On the other hand, the decomposition of a given spectrogram into individual notes can be accomplished more accurately when the fundamental frequencies of the concurrent sounds are given. Also, if we know the fundamental frequencies of all the underlying notes in the signal, they can constitute very useful information for accurately estimating their onset times and vice versa. Furthermore, as the onset times of notes are usually 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 and tempo. If we know the beat locations of a piece of music, then it is much easier to detect the onset times of notes and vice versa, since the inter-onset times are likely to be multiples or fractions of the beat period.

Simultaneous estimation is generally preferable when several estimation problems are interdependent. Thus, we consider it necessary to introduce a unified model, which could be used to jointly solve the problems of determining the pitch and onset time of each musical note, the rhythm and the overall tempo variation of a piece of music. In this paper, we take a Bayesian approach (a generative model

approach) as in [3–6] to formulate and solve this simultaneous estimation problem.

2. GENERATIVE MODEL OF SPECTROGRAM

2.1 Overview

Motivated by the above, this paper proposes modeling the generative process of an entire spectrogram of a piece of music by formulating the following three sub-processes and combining them into one process: (1) the sub-process by which the tempo curve of a piece of music is generated, (2) the sub-process by which a set of note onset positions (in terms of the relative time) is generated based on a 2-dimensional tree structure representation of music, and (3) the sub-process by which a music spectrogram is generated according to the tempo curve generated by sub-process 1 and the set of note onset positions generated by sub-process 2. In the following, we model sub-process 1 in 2.2, sub-process 2 in 2.3 and sub-process 3 in 2.4, respectively. Our aim is to use this complete generative model to explain how a given spectrogram is generated. The most likely model parameters given the observation would then give a musically likely interpretation of what is actually happening in the spectrogram (*i.e.*, a musical score). To this end, we employ a Bayesian approach to infer the posterior distributions of all the latent parameters. An approximate posterior inference algorithm is derived, which is described in Section 3.

2.2 Sub-process for generating tempo curve

The tempo of a piece of music is not always constant and in most cases it varies gradually over time. If we use a [1] "tick" as a relative time notion, an instantaneous (or local) tempo may be defined as the length of 1 tick in seconds. Now let us use μ_d to denote the real duration (in units of seconds) corresponding to the interval between d and $d+1$ ticks. Thus, μ_d corresponds to the local tempo and so the sequence $\mu_1, \ldots, \mu_D$ can be regarded as the overall tempo curve of a piece of music. One reasonable way to ensure a smooth overall change in tempo is to place a Markov-chain prior distribution over the sequence $\mu_1, \ldots, \mu_D$ that is likely to generate a sequence $\mu_1, \ldots, \mu_D$ such that $\mu_1 \simeq \mu_2, \mu_2 \simeq \mu_3, \ldots, \mu_{D-1} \simeq \mu_D$. Here, we assume a Gaussian-chain prior for convenience:

$$\mu_d | \mu_{d-1} \sim \mathcal{N}(\mu_d; \mu_{d-1}, (\sigma^{\mu})^2) \ \ (d = 2, \ldots, D), \quad (1)$$

where $\mathcal{N}(x; \mu, \sigma) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. If we use ψ_d to denote the absolute time (in units of seconds) corresponding to d ticks, ψ_d can thus be written as $\psi_d = \psi_{d-1} + \mu_{d-1}$, which plays the role of mapping a relative time in units of ticks (integer) to an absolute time in units of seconds (continuous value).

2.3 Sub-process for generating note onset positions

Here we describe the generative model of the set of some number R of note onset positions $S_1, \ldots, S_R$ (in units of ticks). Most people would probably agree that music has a 2-dimensional hierarchical structure. Frequent motifs, phrases or melodic themes consist of a hierarchy that can be described as time-span trees. In addition, polyphony often consists of multiple independent voices. That is, we can assume that music consists of a time-spanning tree structure and a synchronizing structure of multiple events at several levels of a hierarchy. We would like to describe this 2-dimensional tree structure representation of music in the form of a generative model. This can be accomplished by introducing a generative model that is conceptually similar to the one proposed in [6].

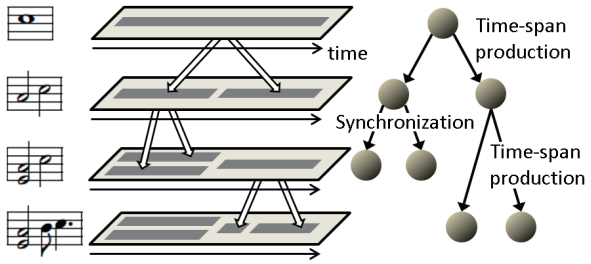


Figure 1. Generative model of a 2-dimensional tree structure representation of musical notes.

Fig. 1 shows an example of the generative process of four musical notes in one bar of 4/4. In this example, a whole note is first split into two consecutive half notes. We call this process "time-spanning." Next, the former half note is copied in the same location, thus resulting in a chord of two half notes. We call this process "synchronization." A chord with an arbitrary number of notes can thus be generated by successively employing this type of binary production. Finally, the latter half note is split into a quaver and a dotted quarter note via the time-spanning process. This kind of generative process can be modeled by extending the idea of the probabilistic context-free grammar (PCFG) [7]. For simplicity, this paper focuses only on Chomsky normal form grammars, which consist of only two types of rules: emissions and binary productions. A PCFG is a pair consisting of a context-free grammar (a set of symbols and productions of the form $A \rightarrow BC$ or $A \rightarrow w$, where A, B, and C are called "nonterminal symbols" and w is called a "terminal symbol") and production probabilities, and defines a probability distribution over the trees of symbols. The parameters of each symbol consist of (1) a distribution over rule types, (2) an emission distribution over terminal symbols, and (3) a binary production over pairs of symbols.

To describe the generative process shown in Fig. 1, we must introduce an extension of PCFG. As we explain later, we explicitly incorporate a process of stochastically choosing either "time-spanning" or "synchronization" in the binary production process. Fig. 2 defines the proposed generative process of the set of the onset positions of some number R of musical notes. In our model, each node n of the parse tree corresponds to one musical note (with no pitch information) and a pair consisting of the onset position S_n and duration L_n of that note is considered to be a nonterminal symbol. We first draw a "switching" distribution (namely, a Bernoulli distribution) ϕ^{T} over the two rule types {EMISSION, BINARY-PRODUCTION} from a Beta distribution. Next, we draw another "switching" dis-

[1] Tick is a relative measure of time represented by the number of discrete divisions a quarter note has been split into. So, if we consider 16 divisions per quarter note, for instance, the duration of 40 ticks corresponds to two-and-a-half beats.

Draw rule probabilities:

$\phi^T \sim \mathrm{Beta}(\phi^{\mathrm{T}}; 1, \beta^{\mathrm{T}})$

[Probability of choosing either of two rule types]

$\phi^N \sim \mathrm{Beta}(\phi^N; 1, \beta^N)$

[Probability of choosing either of two binary-production types]

For each duration l:

$\boldsymbol{\phi}_l^B \sim \mathrm{Dirichlet}(\boldsymbol{\phi}_l^B; \boldsymbol{\beta}_l^B)$

[Probability of position at which segment of length l is split]

For each node n in the parse tree:

$b_n \sim \mathrm{Bernoulli}(b_n; \phi^{\mathrm{T}})$

[Choose either EMISSION or BINARY-PRODUCTION]

If $b_n =$ EMISSION

$S_r \sim \delta_{S_r, S_n}, \;\; L_r \sim \delta_{L_r, L_n}$

[Emit terminal symbol]

If $b_n =$ BINARY-PRODUCTION

$\rho_n \sim \mathrm{Bernoulli}(\rho_n; \phi^{\mathrm{N}})$

[Choose either SYNCHRONIZATION or TIME-SPANNING]

If $\rho_n =$ SYNCHRONIZATION

$S_{n_1} \sim \delta_{S_{n_1}, S_n}, \;\; S_{n_2} \sim \delta_{S_{n_2}, S_n}$

$L_{n_1} \sim \delta_{L_{n_1}, L_n}, \;\; L_{n_2} \sim \delta_{L_{n_2}, L_n}$

[Produce two copies of note n]

If $\rho_n =$ TIME-SPANNING

$S_{n_1} \sim \delta_{S_{n_1}, S_n}, \;\; S_{n_2} \sim \delta_{S_{n_2}, S_n + L_{n_1}}$

$L_{n_1} \sim \delta_{L_{n_1}, L_n - L_{n_2}}$

$L_{n_2} \sim \mathrm{Discrete}(L_{n_2}; \boldsymbol{\phi}_{L_n}^{\mathrm{B}})$

[Split note n into two consecutive notes n_1 and n_2]

Figure 2. The probabilistic specification of the present generative model of a 2-dimensional tree structure representation of musical notes. δ denotes Kronecker's delta. Thus, $x \sim \delta_{x,y}$ means $x = y$ (with probability 1). $\mathrm{Bernoulli}(x; y)$ and $\mathrm{Beta}(y; \boldsymbol{z})$ are defined as $\mathrm{Bernoulli}(x; y) = y^x(1-y)^{1-x}$ and $\mathrm{Beta}(y; \boldsymbol{z}) \propto y^{z_1-1}(1-y)^{z_2-1}$, where $x \in \{0, 1\}$, $0 \leq y \leq 1$ and $\boldsymbol{z} = (z_1, z_2)$, respectively. $\mathrm{Discrete}(x; \boldsymbol{y})$ and $\mathrm{Dirichlet}(\boldsymbol{y}; \boldsymbol{z})$ are defined as $\mathrm{Discrete}(x; \boldsymbol{y}) = y_x$ and $\mathrm{Dirichlet}(\boldsymbol{y}; \boldsymbol{z}) \propto \prod_i y_i^{z_i - 1}$ where $\boldsymbol{y} = (y_1, \ldots, y_I)$ with $y_1 + \cdots + y_I = 1$ and $\boldsymbol{z} = (z_1, \ldots, z_I)$, respectively.

tribution ϕ^{N} over the two binary-production types {TIME-SPANNING, SYNCHRONIZATION} similarly from a Beta distribution. Finally, we generate a discrete distribution $\boldsymbol{\phi}_l^{\mathrm{B}} = (\phi_{l,1}^{\mathrm{B}}, \ldots, \phi_{l,l}^{\mathrm{B}})$ over the position l' at which the segment of duration l is split when BINARY-PRODUCTION is chosen. The shapes of all the Beta distributions and the Dirichlet distribution in our model are governed by concentration hyperparameters: β^{T}, β^{N} and $\boldsymbol{\beta}_1^{\mathrm{B}}, \ldots, \boldsymbol{\beta}_D^{\mathrm{B}}$.

Given a grammar, we generate a parse tree in the following manner: start with a root node that has the designated root symbol, $S_{\mathrm{Root}} = 0$ and $L_{\mathrm{Root}} = D$ where D denotes the overall length of a piece of music in ticks. For each nonterminal node n, we first choose a rule type b_n using ϕ^{T}. If $b_n =$ EMISSION, we produce a terminal symbol S_r with the value of S_n, namely the onset position of note r. If $b_n =$ BINARY-PRODUCTION, we then choose a binary-production type ρ_n using ϕ^{N}. If $\rho_n =$ SYNCHRONIZATION, we produce two nonterminal children n_1 and n_2 such that $S_{n_1} = S_{n_2} = S_n$, $L_{n_1} = L_{n_2} = L_n$. This means that the notes of the child nodes have exactly the same onset and duration. If $\rho_n =$ TIME-SPANNING, we produce two nonterminal children n_1 and n_2 with $S_{n_1} = S_n$, $L_{n_1} = L_n - L_{n_2}$, $S_{n_2} = S_n + L_{n_1}$ where L_{n_2} is drawn from a discrete distribution $\boldsymbol{\phi}_{L_n}^{\mathrm{B}}$. L_{n_2} corresponds to the position at which the segment of duration L_n is divided. We apply the procedure recursively to any nonterminal children and finally obtain a sequence $S_1, \ldots, S_R$ corresponding to the onset positions of R musical notes.

None of the notes r yet contains pitch information. We assign a pitch index κ_r to each note r in the same way as an ordinary cluster assignment process:

$$\boldsymbol{\phi}_r^{\mathrm{K}} \sim \mathrm{Dirichlet}(\boldsymbol{\phi}_r^{\mathrm{K}}; \boldsymbol{\alpha}^{\mathrm{K}}), \tag{2}$$

$$\kappa_r \sim \mathrm{Discrete}(\kappa_r; \boldsymbol{\phi}_r^{\mathrm{K}}), \tag{3}$$

where $\mathrm{Discrete}(x; \boldsymbol{y}) = y_x$ (where $\boldsymbol{y} = (y_1, \ldots, y_I)$ with $y_1 + \cdots + y_I = 1$) and $\mathrm{Dirichlet}(\boldsymbol{y}; \boldsymbol{z}) \propto \prod_i y_i^{z_i - 1}$ (where $\boldsymbol{z} = (z_1, \ldots, z_I)$). The k-th element of $\boldsymbol{\phi}_r^{\mathrm{K}}$ defines how likely each pitch index is to be chosen. It should be noted here that the generative processes of S_r and κ_r should not be considered independently, since harmony and rhythm are in general interdependent of each other. An interesting direction for future work is the joint modeling of these two generative processes.

2.4 Sub-process for generating spectrogram

We now turn to describing the sub-process by which a music spectrogram is generated. Here, we consider that a music spectrogram is generated according to the tempo curve and the set of note onset positions, that have been generated by the sub-processes described in 2.2 and 2.3. To model a spectrogram of a musical audio signal, we make the following assumptions about musical notes:

(A1) Each musical note has a static spectral profile characterized by a particular pitch.

(A2) The magnitude spectrum of music at a certain time instant is represented by a superposition of the spectra of multiple musical notes.

(A3) The power of each musical note varies smoothly in time in the interval between the onset and offset.

From assumption (A1), a magnitude spectrogram of a musical note r can be described as

$$X_{\omega,t} = \sum_{r=1}^{R} H_{\omega,\kappa_r} W_{r,t}, \tag{4}$$

where ω and t are frequency and time indices, respectively. A set consisting of $H_{1,k}, \ldots, H_{\Omega,k} \geq 0$ represents the static spectrum of the k-th pitch and so a set consisting of $H_{1,\kappa_r}, \ldots, H_{\Omega,\kappa_r}$ signifies the spectrum of note r. $W_{r,t} \geq 0$ denotes the power of note r at time t. As the assumptions (A1) and (A2) do not always hold exactly in reality, an actual music spectrogram $Y_{\omega,t}$ may diverge from the "ideal model" $X_{\omega,t}$ to some extent. One way to simplify this kind of deviation process is to assume a probability distribution

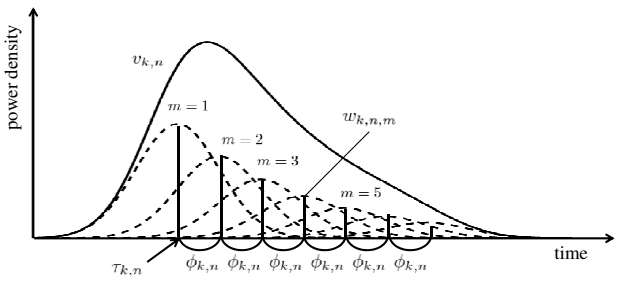


Figure 3. Power envelope $W_{r,t}$ of musical note r.

on $Y_{\omega,t}$ with the expected value of $X_{\omega,t}$. Here, we assume that $Y_{\omega,t}$ follows a Poisson distribution with mean $X_{\omega,t}$:

$$Y_{\omega,t} \sim \mathrm{Poisson}(Y_{\omega,t}; X_{\omega,t}), \tag{5}$$

where $\mathrm{Poisson}(y; x) = x^y e^{-x}/y!$. It should be noted that the maximization of the Poisson likelihood with respect to $X_{\omega,t}$ amounts to optimally fitting $X_{\omega,t}$ to $Y_{\omega,t}$ by using the I-divergence as the fitting criterion [3, 8]. To avoid any indeterminacy in the scaling of H_{ω,κ_r} and $W_{r,t}$, we assume

$$\sum_{\omega} H_{\omega,k} = 1 \quad (k = 1, \ldots, K). \tag{6}$$

Each spectral profile $H_{\omega,k}$ must have the harmonic structure of a particular pitch. One way of ensuring this is to assume a prior distribution over $H_{\omega,k}$ so that it is likely to generate a spectrum with a certain harmonic structure of the k-th pitch. Here, we choose to place a Gamma prior over $H_{\omega,k}$, namely

$$H_{\omega,k} \sim \mathrm{Gamma}(H_{\omega,k}; \gamma \bar{H}_{\omega,k} + 1, \beta), \tag{7}$$

where $\mathrm{Gamma}(x; a, b) \propto x^{a-1} e^{-bx}$. The mode of this prior distribution is given by $\bar{H}_{\omega,k}$, which should be defined such that it corresponds to the most likely spectral profile for the k-th pitch. β determines the peakiness of the density around the mode.

To incorporate assumption (A3) into $W_{r,t}$, we propose describing $W_{r,t}$ using a parametric model expressed as a sum of Gaussians [8] (Fig. 3):

$$W_{r,t} = \sum_{m=1}^{M} G_{r,m,t}, \tag{8}$$
$$G_{r,m,t} = \frac{w_r u_{r,m}}{\sqrt{2\pi}\varphi} e^{-(t-(m-1)\varphi-\tau_r)^2/2\varphi^2},$$

where w_r is the total energy of note r, and τ_r is the center of the first Gaussian, which can be considered the onset time of note r (in seconds). The centers of the Gaussians are constrained so that they are equally spaced with the distance φ, which is equal to the "standard deviation" of all the Gaussians. $u_{r,1}, \ldots, u_{r,M}$ are weights associated with the M Gaussians, which determine the overall shape of the power envelope. To avoid any indeterminacy in the scaling of w_r and $u_{r,m}$, we assume

$$\forall_r : \sum_{m=1}^{M} u_{r,m} = 1. \tag{9}$$

The number of consecutive Gaussians with non-zero weights corresponds to the duration of the note, which we hope to infer automatically from an observed spectrogram. To this end, we choose to use a stick-breaking representation [9] to describe the generative process of $u_{r,1}, \ldots, u_{r,M}$:

$$V_{r,m} \sim \mathrm{Beta}(V_{r,m}; 1, \beta_r^{\mathrm{V}}) \tag{10}$$
$$u_{r,m} = V_{r,m} \prod_{m'=1}^{m-1} (1 - V_{r,m'}), \tag{11}$$

which contributes to sparsifying the Gaussian weights.

Now, recall that the onset position S_r (in ticks) of note r is assumed to have been generated via the generative process described in 2.3. The onset position τ_r of note r should thus be placed near the absolute time into which S_r is converted. Recall also that ψ_d, which can be considered a function that takes a relative time d as an input and returns the corresponding absolute time as an output, is also assumed to have been generated (via the generative process described in 2.2). Given S_r and ψ_d, we find it convenient to write the generative process of τ_r as

$$\tau_r \sim \mathcal{N}(\tau_r; \psi_{S_r}, (\sigma^\tau)^2). \tag{12}$$

2.5 Expansion of generative process

We can describe an expanded version of the generative process of $Y_{\omega,t}$ as

$$C_{r,m,\omega,t} \sim \mathrm{Poisson}(C_{r,m,\omega,t}; H_{\omega,\kappa_r} G_{r,m,\omega,t})$$
$$Y_{\omega,t} \sim \delta\big(Y_{\omega,t} - \sum_{r,m} C_{r,m,\omega,t}\big), \tag{13}$$

by introducing an auxiliary variable $C_{r,m,\omega,t}$. For convenience of analysis, we use this generative process instead of (5) in the following. Note that it can be readily verified that marginalizing out $C_{r,m,\omega,t}$ reduces (13) to (5).

3. APPROXIMATE POSTERIOR INFERENCE

3.1 Variational Bayesian approach

In this section, we describe an approximate posterior inference algorithm for our generative model based on variational inference. The random variables of interest in our model are

$H = \{H_{\omega,k}\}_{\omega,k}$: spectrum of pitch k,
$w = \{w_r\}_r$: total energy of note r,
$V = \{V_{r,m}\}_{r,m}$: shape of power envelope of note r,
$\tau = \{\tau_r\}_r$: onset time (sec) of note r,
$\kappa = \{\kappa_r\}_r$: pitch index assigned to note r,
$\psi = \{\psi_d\}_d$: absolute time corresponding to d ticks,
$\mu = \{\mu_d\}_d$: local tempo between d and $d+1$ ticks,
$S = \{S_r\}_r$: onset position of note r (in ticks),
$L = \{L_r\}_r$: duration of note r (in ticks), and
ϕ^{B}, ϕ^{T}, ϕ^{N}, ϕ^{K} : rule probabilities,

which we denote as Θ. Our goal is to compute the posterior $p(\Theta, C|Y)$ where $Y = \{Y_{\omega,t}\}$ and $C = \{C_{r,m,\omega,t}\}$ are sets consisting of observed magnitude spectra and auxiliary variables, respectively. By using the conditional distributions defined in 2.2, 2.3, 2.4, and 2.5, we can write the

joint distribution $p(Y, \Theta, C)$ as

$$\begin{aligned} &p(Y, H, w, V, \tau, \kappa, \psi, \mu, S, L, \boldsymbol{\phi}^B, \phi^T, \phi^N, \boldsymbol{\phi}^K, C) \\ &= p(Y|C)p(C|H, w, V, \tau, \kappa)p(H)p(V)p(w) \\ &\quad p(\tau|\psi, S)p(\psi|\mu)p(\mu)p(\kappa|\boldsymbol{\phi}^K)p(\boldsymbol{\phi}^K) \\ &\quad p(S, L|\boldsymbol{\phi}^B, \phi^T, \phi^N)p(\boldsymbol{\phi}^B)p(\phi^T)p(\phi^N), \end{aligned} \tag{14}$$

but to obtain the exact posterior $p(\Theta, C|Y)$, we need to compute $p(Y)$, which involves many intractable integrals.

We can express this posterior variationally as the solution to an optimization problem:

$$\underset{q \in \mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}(q(\Theta, C) \| p(\Theta, C|Y)), \tag{15}$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence between its two arguments. Indeed, if we let $\mathcal{Q}$ be the family of all distributions over Θ and C, the solution to the optimization problem is the exact posterior $p(\Theta, C|Y)$, since KL divergence is minimized exactly when its two arguments are equal. Of course, solving this optimization problem is just as intractable as directly computing the posterior. Although it may appear that no progress has been made, having a variational formulation allows us to consider tractable choices of $\mathcal{Q}$ in order to obtain principled approximate solutions.

For our model, we define the set of approximate distributions $\mathcal{Q}$ as those that factor as follows:

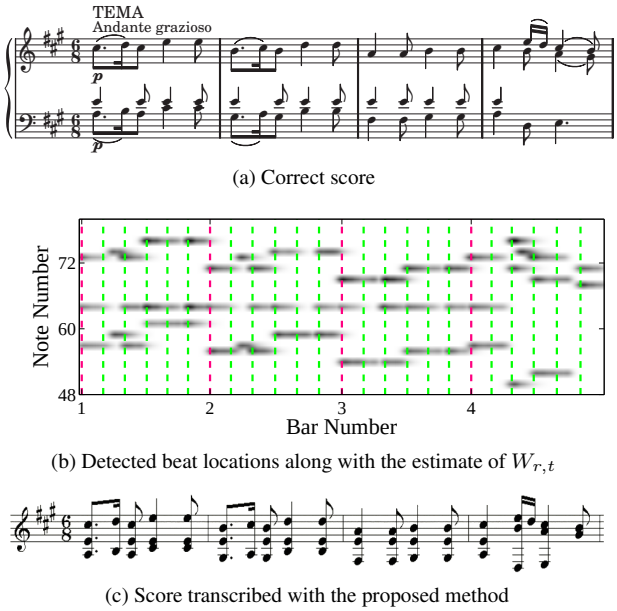
$$\begin{aligned} \mathcal{Q} = \big\{ q : q(C)q(H)q(w)q(V)q(\tau, \psi, \mu)q(\kappa) \\ q(S, L)q(\phi^{\mathrm{K}})q(\phi^{\mathrm{B}})q(\phi^{\mathrm{T}})q(\phi^{\mathrm{N}}) \big\}. \end{aligned} \tag{16}$$

We admit that this is a strong assumption. Its validity and how it affects the parameter inference result must be investigated in the future.

3.2 Coordinate ascent

We now present an algorithm for solving the optimization problem described in (15) and (16). Unfortunately, the optimization problem is non-convex, and it is intractable to find the global optimum. However, we can use a simple coordinate ascent algorithm to find a local optimum. The algorithm optimizes one factor in the mean-field approximation of the posterior at a time while fixing all the other factors. The mean-field update equations for the variational distributions are given in the following form:

$$q(\boldsymbol{C}_{\omega,t}) = \mathrm{Multinomial}(\boldsymbol{C}_{\omega,t}; Y_{\omega,t}, \boldsymbol{f}^C_{\omega,t}), \tag{17}$$
$$q(H_{\omega,k}) = \mathrm{Gamma}(H_{\omega,k}; \xi^H_{\omega,k}, \zeta^H_{\omega,k}), \tag{18}$$
$$q(w_r) = \mathrm{Gamma}(w_r; \xi^w_r, \zeta^w_r), \tag{19}$$
$$q(V_{r,m}) = \mathrm{Beta}(V_{r,m}; \xi^V_{r,m}, \zeta^V_{r,m}), \tag{20}$$
$$q(\boldsymbol{\tau}, \boldsymbol{\psi}, \boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\chi}; \boldsymbol{\xi}^\chi, \boldsymbol{\zeta}^\chi), \tag{21}$$
$$q(\kappa_r) = \mathrm{Discrete}(\kappa_r; \boldsymbol{f}^\kappa_r), \tag{22}$$
$$q(\boldsymbol{\phi}^K_r) = \mathrm{Dirichlet}(\boldsymbol{\phi}^K_r; \boldsymbol{\xi}^K_r), \tag{23}$$
$$q(S_r, L_r) = \mathrm{Discrete}(S_r, L_r; \boldsymbol{f}^{SL}_r), \tag{24}$$
$$q(\boldsymbol{\phi}^B_l) = \mathrm{Dirichlet}(\boldsymbol{\phi}^B_l; \boldsymbol{\xi}^B_l), \tag{25}$$
$$q(\phi^T) = \mathrm{Beta}(\phi^T; \xi^T, \zeta^T), \tag{26}$$



(a) Correct score

(b) Detected beat locations along with the estimate of $W_{r,t}$

(c) Score transcribed with the proposed method

Figure 4. Transcription result obtained with the proposed method applied to Morzart: Piano Sonata No. 11 in A major, K. 331/300i under the situation where $\tau_1, \ldots, \tau_R$ are given. In (b), the red and green lines indicate the estimates of bar lines and the positions of beat locations obtained with the present method, respectively.

$$q(\phi^N) = \mathrm{Beta}(\phi^N; \xi^N, \zeta^N), \tag{27}$$

where

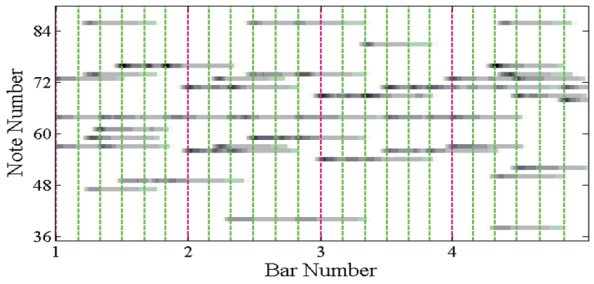
$$\boldsymbol{\chi} = \begin{bmatrix} \boldsymbol{\tau} \\ \boldsymbol{\psi} \\ \boldsymbol{\mu} \end{bmatrix}, \quad \boldsymbol{\xi}^\chi = \begin{bmatrix} \boldsymbol{\eta}^\tau \\ \boldsymbol{\eta}^\psi \\ \boldsymbol{\eta}^\mu \end{bmatrix}, \quad \boldsymbol{\zeta}^\chi = \begin{bmatrix} \boldsymbol{\nu}^\tau & \boldsymbol{\nu}^{\tau\psi} & \boldsymbol{\nu}^{\tau\mu} \\ \boldsymbol{\nu}^{\tau\psi} & \boldsymbol{\nu}^\psi & \boldsymbol{\nu}^{\psi\mu} \\ \boldsymbol{\nu}^{\tau\mu} & \boldsymbol{\nu}^{\psi\mu} & \boldsymbol{\nu}^\mu \end{bmatrix}.$$

(25)–(27) are performed only when we want to learn the rule probabilities. (24)–(27) can be updated using the inside-outside algorithm. The update formulas of the variational parameters are all given in analytical form, but they are omitted here owing to space limitations.

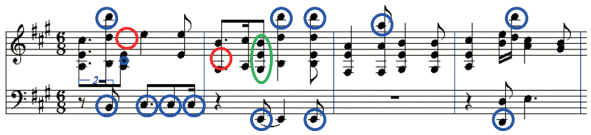
4. EXPERIMENTAL RESULTS

We now present experimental results obtained with our proposed model. We first conducted a preliminary experiment to confirm that our model can transcribe a score (appropriately estimate the note values of musical notes, beat locations, and the tempo of a music piece) when the onset times of all the musical notes (namely, τ_r's) are given. We then show an example of transcription results obtained using the complete model directly from an audio spectrogram.

For the first experiment, we used a few piano recordings (RWC-MDB-C-2001 No. 26, 27, 30) excerpted from the RWC music database [12]. The data were the first 10 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, 30 Hz and 12 cents, respectively. In this experiment, all the values $\tau_1, \ldots, \tau_R$ were given manually. The hyperparameters and initial parameters were set at $K = 74, M = 40, \varphi = 3, \alpha^H_{\omega,k} = \beta^H_{\omega,k}\bar{H}_{\omega,k}+1, \beta^H_{\omega,k} = 500, \alpha^w_r =$



(a) Detected beat locations along with the estimate of $W_{r,t}$



(b) Score transcribed with the proposed method

Figure 5. Transcription result obtained with the proposed method applied to Morzart: Piano Sonata No. 11 in A major, K. 331/300i. In (a), the red and green lines indicate the estimates of bar lines and the positions of beat locations obtained with the present method, respectively. In (b), the red, green and blue circles indicate the deletion errors, pitch errors and octave errors, respectively.

$\beta_r^w = 0, \beta_{r,m}^V = 10e^{-m/8}/\sum_{m'} e^{-m'/8}, \sigma^\tau = 2, \sigma^\psi = 1, \sigma^\mu = 0.5, \alpha_{r,k} = 2, \beta^T = 1, \beta^N = 2$. The initial values of $H_{\omega,k}$ and $\bar{H}_{\omega,k}$ were set at the value obtained with the non-netaive matrix factorization [13] applied to the magnitude spectrogram of the piano excerpts from the RWC musical instrument sound database [11]. We set the resolution of the relative time at 4 ticks per quarter note. D and the initial values of ψ_d were set at the values obtained with [10]. The algorithm was run for 10 iterations. After convergence, we took the expected values of the posteriors and regarded them as the parameter estimates.

Fig. 4 shows an example of the score we obtained when we applied the present method to Mozart's Sonata (RWC-MDB-C-2001 No. 26). As can be seen from this example, the note values and the beat locations were appropriately estimated. We also confirmed that reasonably good results were obtained for other recordings such as Chopin's Nocturne No. 2 in E♭-maj, Op. 9 (RWC-MDB-C-2001 No. 30).

For the second experiment, we applied our method without providing any information about τ. The experimental conditions were the same as above except that we assumed that τ was unknown. Fig. 5 shows an example of the estimates of $W_{r,t}$ (namely, the power envelope of note r) and the score obtained with the present method applied to the same data in Fig. 5. The result showed that many octave errors had occurred. This kind of error often occurs when there is a mismatch between a spectrum model and an actual spectrum. The validity of the assumptions we have made about the spectra of musical sounds in 2.4 must be carefully examined in the future.

5. CONCLUSION

This paper proposed a Bayesian model for automatic music transcription. Automatic music transcription involves several interdependent subproblems: multiple fundamental frequency estimation, onset detection, and rhythm/tempo recognition. To circumvent the chicken-and-egg problem, we modeled the generative process of an entire music spectrogram by combining the sub-process by which a musically natural tempo curve is generated, the sub-process by which a set of note onset positions is generated based on a 2-dimensional tree structure representation of music, and the sub-process by which a music spectrogram is generated according to the tempo curve and the note onset positions. Thanks to this combined generative model, the present method performs note extraction and structure estimation simultaneously and thus an optimal solution is obtained within a unified framework. We described some of the transcription results obtained with the present method.

6. REFERENCES

[1] 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. ICASSP2007*, Vol. 1, pp. 65–68, 2007.

[2] K. Ochiai, H. Kameoka, and S. Sagayama, "Explicit beat structure modeling for non-negative matrix factorization-based multipitch analysis," in *Proc. ICASSP2012*, pp. 133–136, 2012.

[3] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," Technical Report CUED/F-INFENG/TR.609, University of Cambridge, 2008.

[4] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. ICML2010*, pp. 439–446.

[5] K. Yoshii, and M. Goto, "A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation," *IEEE Trans. Audio, Speech, Language Process.*, Vol. 20, No. 3, pp. 717–730, 2012.

[6] M. Nakano, Y. Ohishi, H. Kameoka, R. Mukai, and K. Kashino, "Bayesian nonparametric music parser," in *Proc. ICASSP2012*, pp. 461–464, 2012.

[7] P. Liang, S. Petrov, M. I. Jordan, and D. Klein, "The infinite PCFG using hierarchical Dirichlet processes," in *EMNLP2007*, pp. 688–697.

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

[9] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, Vol. 4, pp. 639–650, 1994.

[10] D. P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, 36(1):51–60, 2007.

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

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

[13] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for music transcription," in *Proc. WASPAA2003*, pp. 177–180, 2003.