ABSTRACT
This paper proposes a novel representation of music that can be used for similarity-based music information retrieval, and also presents a method that converts an input polyphonic audio signal to the proposed representation. The representation involves a 2-dimensional tree structure, where each node encodes the musical note and the dimensions correspond to the time and simultaneous multiple notes, respectively. Since the temporal structure and the synchrony of simultaneous events are both essential in music, our representation reflects them explicitly. In the conventional approaches to music representation from audio, note extraction is usually performed prior to structure analysis, but accurate note extraction has been a difficult task. In the proposed method, note extraction and structure estimation is performed simultaneously and thus the optimal solution is obtained with a unified inference procedure. That is, we propose an extended 2-dimensional infinite probabilistic context-free grammar and a sparse factor model for spectrogram analysis. An efficient inference algorithm, based on Markov chain Monte Carlo sampling and dynamic programming, is presented. The experimental results show the effectiveness of the proposed approach.

Index Terms— infinite probabilistic context-free grammar (infinite PCFG), nonnegative matrix factorization (NMF), Markov chain Monte Carlo (MCMC), hierarchical Dirichlet process (HDP)

1. INTRODUCTION
The analysis of musical audio signals has been a very active area of research. One of the tasks most frequently addressed in the field has been automatic music transcription, where the music audio is represented as a score [1, 2]. However, transcription from polyphonic music audio signals has continued to be a difficult task. On the other hand, it is widely known that music can be perceived within a hierarchical structure over time, namely, frequent motifs, phrases, melodic themes, or larger sections such as verses or chorus parts, where dominant elements contain subordinate elements. Since such a structure and multiple note events are both essential elements embedded in music, these problems should be addressed simultaneously.

This paper proposes a novel hierarchical representation of music, and a parser, which is a method of obtaining the proposed representation from an input polyphonic music audio signal. The representation involves a 2-dimensional tree-structure. Each node encodes the musical note, and the dimensions correspond to the time and simultaneous multiple notes, respectively. Applications of the parser and the representation we have in mind include content-based music information retrieval systems. For example, even when the tempo, style, and instrumentation of the songs vary in a cover song identification task, the trees directly give us a clue to the frequent motifs or melodic themes. Moreover, the representation can potentially be used to accelerate the music audio search because efficient search techniques can be applied to the tree structure.

Conventionally, the generative theory of tonal music (GTTM) [3] is a well-known approach with which to understand musical intuitions [4]. “Time-span tree” [3] is used to represent a hierarchical structure of events. While their methods are mainly based on the empirical rules for constructing trees, we aim to obtain production rules even in a probabilistic framework. Moreover, we also discuss a tree-structured representation for polyphonic scores, although a “time-span tree” has generally been applied to monophonic scores.

Based on a Bayesian nonparametric framework, the proposed parser runs with little previous knowledge, and simultaneously optimizes both the estimated structure and the estimated notes through a unified inference algorithm. As described in the following sections, the method comprises an extended 2-dimensional infinite probabilistic context-free grammar (PCFG) [5] and a sparse factor model for spectrogram analysis [1, 6], and employs Markov chain Monte Carlo sampling and dynamic programming for inference.

2. BAYESIAN NONPARAMETRIC MUSIC PARSER
Music has a 2-dimensional hierarchical structure. Frequent motifs, phrases or melodic themes consists of a hierarchy, which can be expressed as time-span trees. In addition, polyphony often has multiple independent voices. That is, we can consider that music consists of a time-spanning structure and the synchronization of multiple events at several levels of a hierarchy. We present a Bayesian model of 2-dimensional tree structures as a representation of music.

As shown in Fig. 1, the 2-dimensional tree-structured representation can be regarded as a possible generative model. Note that we have not yet determined the pitch or timbral information for each note (discussed in Section 3). Fig. 1 shows the generative process of 1 bar. A whole note is first divided into two half notes, which are expressed as the time-span production. The former half note is then copied into the same location, which represents a two-note chord. Note that chords require the concept of the “synchronization” of multiple notes. The latter half note is also divided into a quaver and a dotted quarter note. Such processes are also applied
to several levels of hierarchy. For another instance, the whole entity is divided into "verse", "bridge", and "chorus" (time-spanning structure). The "chorus" part includes the counterpoint between two voices (synchronism).

Such a hierarchy can be modeled with an extension of PCFG, analogous to natural language processing. Since we hope that our model will be applicable to all possible music signals, and parsimonious grammars should be automatically learned depending on input data, we use a Bayesian nonparametric approach for modeling all possible syntactic tree structures. Here, we first review the conventional Bayesian nonparametric PCFG, known as the infinite PCFG [5], to build up an understanding of our model. For simplicity, this paper focuses only on Chomsky normal form grammars, which have 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 nonterminal symbols and $w$ is a terminal symbol) and production probabilities, and defines a probability distribution over trees of symbols. The parameters of each symbol consist of (1) a distribution over rule types, (2) an emission distribution over terminal 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. The infinite PCFG has tackled the question of how to find an adequate number of symbols. It is defined as having an infinite number of symbols by a hierarchical Dirichlet process (HDP) prior. We place the Dirichlet process (DP) prior over symbols: $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \sim \text{DP}(\gamma, I)$ where $I$ is a base measure over symbols and $\gamma$ is a concentration parameter. The DP is a distribution over "length" space (\cite{7}).

In the previous section, we constructed a generative model for a 2-dimensional tree structure. It can be used as a prior distribution on a process (DP) prior over symbols:

$$G_k \sim \text{DP}(\gamma, I \times F),$$

where $F$ provides a distribution over "length" space ($\in \mathbb{R}$). $L_0 (\theta = 1, 2, \ldots)$ denotes the length embedded in the $\theta$-th symbol. We can consider $G_0$ to be a probabilistic measure over symbols with latent lengths. As in the infinite PCFG, we draw from the following process:

$$G'_0 = \sum_{i,j} \beta_i \beta_j \delta_{(\phi_i, L_i)} \times \delta_{(\phi_j, L_j)} \sim \text{DP}(\alpha, G'_0),$$

Second, the distributions over binary productions are formulated by scaling the probabilities of the Dirichlet process:

$$w^{(k)}_{i,j} = \text{exp}\left(-\frac{(L_k - L_i - L_j)^2}{\sigma^2}\right), \quad G_k(i,j) \propto w^{(k)}_{i,j} G'_k(i,j).$$

Intuitively, $w^{(k)}_{i,j}$ represents the similarity between the parent’s length $L_k$ and the sum of the children’s $L_i + L_j$. With this, we can explicitly give priority to the binary productions maintaining the total length of time on parent-child relationships.

The remaining problem is how to generate the 2-dimensional architecture of trees. As shown in Fig. 1, we can consider each tree node as the dominant region on the time axis, which contains smaller elements. First, each node should have not only “length” but also “onset” to mark its placement. Such onset propagations provide more flexibility for tempo fluctuation. Next, to express a synchrony of multiple notes, we introduce a binary indicator $b_m$ into each tree node (indexed by $m$): $b_m \sim \text{Bernoulli}(a_B)$. When $b_m = 1$, the $m$-th binary production is chosen from $G_k$. Otherwise, a special production related to "synchronization" is selected, which makes a pair of the copy of the parent’s symbol and puts them in the same location as the parent’s. We place the prior $\text{Beta}(1, \chi_2)$ on $a_B$.

We now turn to a constructive representation of the proposed model. Various constructions of DP, HDP, and their extensions have been proposed. The normalized Gamma process representation has a high affinity with our model \cite{7}. We use the following finite approximation as the top-level Dirichlet process: $\beta \sim \text{Dirichlet}(\gamma/K, \ldots, \gamma/K)$ where $K$ works as the truncation level. As $K$ increases, our approximation improves. Each distribution over binary productions on the time axis is generated as a modification of the normalized Gamma process representation of the Dirichlet process:

$$Z^{(k)}_{i,j} \sim \text{Gamma}(\alpha(\beta^T, 1/w^{(k)}_{i,j}), 1) \quad \text{and} \quad G_k \sim \sum_{i,j} \sum_{i',j'} Z^{(k)}_{i',j'} \delta((\phi_i, L_i), (\phi_j, L_j)).$$

Note that the Gamma distribution is parameterized by a shape parameter and a rate (inverse-scale) parameter. The generative process of a tree is based on the following rules:

$$\begin{pmatrix} \text{Child}_{\text{left}}, \text{Child}_{\text{right}} \end{pmatrix} \sim \begin{cases} \text{Gamma}(\alpha(\beta^T, 1/w^{(k)}_{i,j}), 1) \quad (b_m = 1) \\ \delta_{(k,k)} \quad (b_m = 0) \end{cases}$$

$$t_l \sim \delta_i, \quad t_r \sim \begin{cases} \text{Normal}(t_p + \text{Length}_{\text{Child}_{\text{left}}}, \rho^2) \quad (b_m = 1) \\ \delta_{t_p} \quad (b_m = 0) \end{cases}$$

where Child_{\text{left}} and Child_{\text{right}} denote indexes of symbols, $t_l$ and $t_r$ show their onsets, and $t_p$ is the onset of their parent node. If the symbol indexed by $k$ is assigned to the $m$-th node and $b_m = 1$, the weight $Z^{(k)}_{i,j} / \sum_{i',j'} Z^{(k)}_{i',j'}$ of $G_k$ gives the probability that Child_{\text{left}} = $i$ and Child_{\text{right}} = $j$ are chosen.

3. FULL GENERATIVE MODEL FROM PARSING TREE TO MUSIC SPECTROGRAM

In the previous section, we constructed a generative model for a 2-dimensional tree structure. It can be used as a prior distribution on a
Bayesian model for music audio signals. Bayesian hierarchical approaches have the advantage that probabilistic values are inferred in unified frameworks. We apply it to sparse factor models, and in particular Bayesian nonnegative matrix factorization (NMF) [1]. The conventional NMF applied to audio signal analysis is based on a music signal model where the magnitude or power spectrogram \( Y = (Y_{u,t})_{u \in \mathbb{R}^2, t = 1, \ldots, T} \), which is a time frame index, is factorized into nonnegative parameters, spectral bases \( H = (H_{u,n})_{u \in \mathbb{R}^2, n = 1, \ldots, N} \) and time-varying gains \( U = (U_{u,t})_{u \in \mathbb{R}^2, t = 1, \ldots, T} \). A generative model can be written as follows [6]:

\[
Y_{u,t} \approx \sum_n H_{u,n} U_{t,n} \sim \text{Poisson}(H_{u,n} U_{t,n})
\]

where \( C_n \) denotes the \( n \)-th hidden component. This implies that a number of events that have a similar spectral pattern are extracted as one component. (For example, when \( A^2 \) is played three times on a piano in the input audio signal, the three events can be expected to be learned as one component.) In contrast, we want to construct a note-level model that makes use of the latent lengths embedded in the symbols are also sampled by the Metropolis-Hastings algorithm. We use the Normal distribution as the proposal distribution, whose mean is given by the posteriors well. When \( \gamma \) is close to zero, the large value of the inverse scale parameter practically destabilizes the sampling procedure. To avoid numerical issues, we choose instead to use the Normal distribution (truncated to hold nonnegativity) with the same mean equal to the above Gamma distribution, and apply an acceptance/rejection scheme based on the Metropolis-Hastings algorithm.

The latent lengths embedded in the symbols are also sampled by the Metropolis-Hastings algorithm. We use the Normal distribution as the proposal distribution, whose mean is given by \( \gamma \). Where \( \gamma \) is the step size of the gradient descent.

5. EXPERIMENTS

We now present some example that we undertook with the proposed method. Audio data were downmixed to mono and downsampled to 16 kHz. A magnitude spectrogram was computed using the short time Fourier transform with a 32 ms long Hanning window and a 16 kHz. As discussed in Section 3, adequate spectral bases and temporal envelopes were trained by using certain state-of-the-art NMF techniques, namely, harmonic constraints for spectral bases [2], different prior distributions to the tonal and percussive signals [9], and Bayesian nonparametrics [1]. We set the hyperparameters as follows: \( \alpha = \gamma = 1, \beta = 1, \phi = \sqrt{2}, \sigma = 1, \eta = 0.1, \chi_1 = 3, \chi_2 = 2, \lambda = \Omega / \sum_{s \in \mathbb{R}^2} Y_{u,t} \). For the base measures of HDP, we set \( I \) as the uniform distribution and \( F \) as the non-informative Gamma distribution.

For the first experiment, we used two segments (bars 2-4 and bars 8-10) extracted manually from the classic song (RWC-MDB-C-2001 No. 24A) [10]. The observation times for each segment were the same length. We applied the proposed algorithm to them (two parsing trees and the shared parameters) with truncation \( N=40 \).
Fig. 2. Piano roll as the ground truth of note events (left) and estimated note events \( \sum_{n} V_{n} O_{n,t-r_{n}} \) (right). We trained 13 spectral bases and 5 envelope patterns in advance. The proposed method captured the nearly adequate number of notes and their pitch information.

Fig. 3. Two examples of parsing trees corresponding to bars 2-4 of RWC-MDB-C-2001 No. 24A (shown in Fig. 2). Each row indicates the region (onset time and length) of each node of the estimated parsing tree on the time axis. Since 2-dimensional tree structures are too complicated to draw, parent-child relationships are not explicitly presented. For instance, with the left sample, the 28-th node is divided into the 36 and 37-th nodes using “synchrony” production. As for the right sample, the 32-nd and 33-rd nodes are generated from the 21-st node using “time-spanning” production.

Fig. 4. Piano roll as the ground truth of note events (top) and estimated tree (bottom).

6. DISCUSSION

We proposed a parser for music signals and the resulting music representation. It is based on Bayesian nonparametric sparse factor analysis and PCFG. We also presented an efficient inference algorithm using MCMC and dynamic programming. Our experiments showed that the proposed method successfully captured the multiple hierarchical structures of music signals. Concerning computational costs, for the first experiment, each MCMC iteration requires approximately ten seconds with 2.5 GHz CPU, in non-optimized Matlab\textsuperscript{TM}. In the future, more sophisticated inference methods will be considered, such as collapsed sampling [8], slice sampler [11], and retrospective sampling [12].

7. REFERENCES