BAYESIAN NONPARAMETRIC MUSIC PARSER

Masahiro Nakano, Yasunori Ohishi, Hirokazu Kameoka, Ryo Mukai, Kunio Kashino

NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan
nakano.masahiro@lab.ntt.co.jp

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.

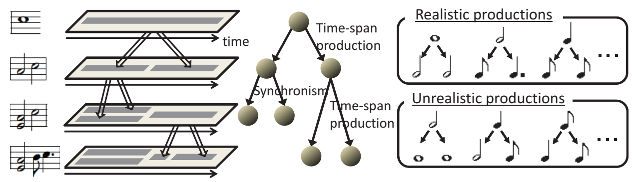

Fig. 1. Generative process of a 2-dimensional hierarchical tree (left) and time-spanning productions (right).

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 "chrous" (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 $A \to B\ C$ or $A \to w$, 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, 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: $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \sim \mathrm{DP}(\gamma, I)$ where I is a base measure over symbols and γ is a concentration parameter of DP. $\boldsymbol{\beta}$ shows symbol probabilities, and ϕ_k $(k = 1, 2, \dots)$ are their atoms. The binary production distributions are drawn from a DP centered on $\boldsymbol{\beta}\boldsymbol{\beta}^T$: $G'_k \sim \mathrm{DP}(\alpha, G'_0)$ $(k = 1, 2, \dots)$, where $G'_0 = \sum_{i,j} \beta_i \beta_j \delta_{(\phi_i, \phi_j)}$. Each G'_k can be intuitively regarded as the infinite-dimensional multinomial distribution over children (the pairs of an unlimited number of symbols) whose parent is indexed by k. All G'_k $(k = 1, 2, \dots)$ share the atoms (the pairs of symbols) drawn from the top-level DP. The PCFG using HDP thus defines the trees consisting of an infinite number of symbols.

In this paper we develop a 2-dimensional infinite PCFG with "length"-embedded symbols to express 2-dimensional hierarchical structures of music. The key difference is that we introduce biased rules into the infinite PCFG. As shown in Fig. 1, time-span productions divide a parent's region into two children's region on the time axis. For example, a half note can be divided into "crotchet and crotchet", or "semiquaver and dotted quarter note". However, it is strange for a half note to be divided into "triplet and triplet". In other words, each node of the trees has its length on the time axis, and the production rules preferably maintain the total length of time in parent-child relationships (Fig. 1). The following strategy is inspired by the HDP with a correlation structure [7]. Since the standard infinite PCFG inherently cannot model time-span structures, we relate the latent "length" on the time axis to each symbol. The production probabilities are modified by their latent lengths.

A top-level Dirichlet process is first drawn with a product base measure:

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{(\phi_k, L_k)} \sim \mathrm{DP}(\gamma, I \times F) \ , \tag{1}$$

where F provides a distribution over "length" space ($\in \mathbb{R}$). L_θ ($\theta = 1, 2, \dots$) denotes the length embedded in the θ-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:

$$G'_0 = \sum_{i,j} \beta_i \beta_j \delta_{((\phi_i, L_i), (\phi_j, L_j))} \ , \ \ G'_k \sim \mathrm{DP}(\alpha, G'_0) \ . \tag{2}$$

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

$$w_{i,j}^{(k)} = \exp\Big(- \frac{(L_k - L_i - L_j)^2}{\sigma^2} \Big), \ \ G_k(i,j) \propto w_{i,j}^{(k)} G'_k(i,j) \ .$$

Intuitively, $w_{i,j}^{(k)}$ 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) : $b_m \sim \mathrm{Bernoulli}(a_B)$. 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 $\mathrm{Beta}(\chi_1, \chi_2)$ 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 [7]. We use the following finite approximation as the top-level Dirichlet process: $\boldsymbol{\beta} \sim \mathrm{Dirichlet}(\gamma/K, \dots, \gamma/K)$ 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:

$$Z_{i,j}^{(k)} \ \sim \ \mathrm{Gamma}(\alpha(\boldsymbol{\beta}\boldsymbol{\beta}^T)_{i,j}, 1/w_{i,j}^{(k)}) \ , \tag{3}$$

$$G_k \ = \ \sum_{i,j} \frac{Z_{i,j}^{(k)}}{\sum_{i',j'} Z_{i',j'}^{(k)}} \delta_{((\phi_i, L_i), (\phi_j, L_j))} \ . \tag{4}$$

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:

$$\Big(\mathrm{Child_{left}}, \mathrm{Child_{right}}\Big) \mid \mathrm{Parent} = k \sim \begin{cases} \Big(\dots, \frac{Z_{i,j}^{(k)}}{\sum_{i',j'} Z_{i',j'}^{(k)}}, \dots \Big) & (b_m = 1) \\ \delta_{(k,k)} & (b_m = 0) \end{cases} \tag{5}$$

$$t_l \sim \delta_{t_p}, \ \ t_r \sim \begin{cases} \mathrm{Normal}(t_p + L_{\mathrm{Child_{left}}}, \rho^2) & (b_m = 1) \\ \delta_{t_p} & (b_m = 0) \end{cases} \tag{6}$$

where $\mathrm{Child_{left}}$ and $\mathrm{Child_{left}}$ 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 $Z_{i,j}^{(k)} / \sum_{i',j'} Z_{i',j'}^{(k)}$ of G_k gives the probability that $\mathrm{Child_{left}} = i$ and $\mathrm{Child_{left}} = j$ 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 $\boldsymbol{Y} = (Y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0}$, where $\omega = 1,\dots,\Omega$ is a frequency bin index, and $t = 1,\dots,T$ is a time frame index, is factorized into nonnegative parameters, spectral bases $\boldsymbol{H} = (H_{\omega,n})_{\Omega\times N}$ and time-varying gains $\boldsymbol{U} = (U_{n,t})_{N\times T}$: $Y_{\omega,t} \approx \sum_n H_{\omega,n}U_{n,t}$. A generative model can be written as follows [6]: $Y_{\omega,t} = \sum_n C_{\omega,t,n}$, $C_{\omega,t,n} \sim \mathrm{Poisson}(H_{\omega,n}U_{n,t})$, where $\boldsymbol{C}_n$ 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 $A\sharp$ 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 symbolic tree-structured prior.

We assume that each note event consists of one spectral basis, its total gain, a temporal envelope pattern (corresponding to one event), and its onset time. The n-th note event chooses the d_n-th spectral pattern $(H_{\omega,d_n})_{\Omega\times 1}$ by the multinomial distribution over the dictionary of spectral patterns $(H_{\omega,d})_{\Omega\times D}$. Note that though the spectral bases including timbral information may be learned from input data on the same principle as NMF, we, in this paper, train them in advance for simplicity. Similarly, the temporal envelope pattern indexed by e_n, $(O_{e_n,t'})_{1\times T'}$, is chosen from the pre-trained dictionary of temporal envelope patterns $(O_{e,t'})_{E\times T'}$, where T' denotes the temporal size. In this paper, we manually fix the sizes of dictionaries, D and E, and the temporal size of envelope patterns T' at the pre-trained stage. We can consider the following generative process for the n-th event:

$$C_{\omega,t,n} \sim \mathrm{Poisson}(H_{\omega,d_n}V_n O_{e_n,t-\tau_n}) \ , \tag{7}$$

where V_n denotes the gain, and the onset time is represented as τ_n. Each τ_n is generated from the onset time t_n of the corresponding terminal symbol: $\tau_n \sim \mathrm{Normal}(t_n,\phi)$. Instead of finding an adequate number of events in advance, we introduce the Gamma process prior for the overall gain of the corresponding source similarly to [1, 6]: $V_n \sim \mathrm{Gamma}(\eta/N, \eta\lambda)$, where N is the truncated number of events.

4. INFERENCE

Our inference strategy is based on an improvement of combined dynamic programming and Markov chain Monte Carlo (a weak-limit sampler using Metropolis-Hastings within Gibbs sampling). Due to limitations of space, we discuss only the core techniques.

Sampling the parsing trees

Since the parse trees propagate the continuous value "onset times" of symbols, standard dynamic programming is complicated for our model. We thus derive computationally tractable approximations by time discretization. "Onset times" are quantized by the truncation level. (We choose the window width of the short-time Fourier transform.) With this, we can describe the following Gibbs sampler using a dynamic programming algorithm [8]. First, we construct the "inside" table, which is similar to the standard Inside-Outside algorithm:

$$p_{n,n',k,t} \propto \sum_{s,k_l,k_r,t_r} a_B \frac{Z^{(k)}_{k_l,k_r}}{\sum_{i,j} Z^{(k)}_{i,j}} \frac{1}{\sqrt{2\pi}\rho} \exp\Big(-\frac{(t_r - t - L_{k_l})^2}{2\rho^2}\Big) \cdot p_{s,n',k_l,t} p_{n,N-s+2,k_r,t_r} + \sum_s (1-a_B) p_{s,n',k,t} p_{n,N-s+2,k,t}.$$

where $p_{n,n',k,t}$ denotes the sum of the probabilities of all trees whose root node is labeled k, k_l and k_r show $\mathrm{Child}_{\mathrm{left}}$ and $\mathrm{Child}_{\mathrm{left}}$ respectively, the onset time is t, and the terminals are from n'-th to n-th series. Next, we obtain the sampling algorithm based on the following multinomial distribution:

$$p(s,k_l,k_r,t_l,t_r) = \frac{p(k \to (k_l,k_r), t \to (t_l,t_r)) p_{s,n',k_l,t_l} p_{n,N-s+2,k_r,t_r}}{p_{n,n',k,t}} \tag{8}$$

where s is the possible "split" position and

$$p(k \to (k_l,k_r), t \to (t_l,t_r)) = a_B \frac{Z^{(k)}_{k_l,k_r}}{\sum_{i,j} Z^{(k)}_{i,j}} \frac{1}{\sqrt{2\pi}\rho} \exp\Big(-\frac{(t_r - t - L_{k_l})^2}{2\rho^2}\Big) + (1-a_B)\delta(k_l = k, k_r = k)\delta(t_r = t). \tag{9}$$

Sampling the weight $\boldsymbol{Z}$ and length $\boldsymbol{L}$

As for the weight of the Gamma process, we choose the approximated conditional distributions as the proposal distributions. We use a first-order Taylor expansion $-\log \sum_{i,j} Z^{(k)}_{i,j} \geq -\log \zeta^{(k)} - (\sum_{i,j} Z^{(k)}_{i,j}/\zeta^{(k)}) + 1$ on the log posterior. The conditional distribution on $Z^{(k)}_{i,j}$ is approximated by the following distribution:

$$q(Z^{(k)}_{i,j}) = \mathrm{Gamma}\Big(\alpha\beta_i\beta_j + N^{(k)}_{i,j},\ 1/w^{(k)}_{i,j} + N^{(k)}_{i,j}/\zeta^{(k)}\Big) \tag{10}$$

where $N^{(m)}_{i,j}$ denotes the number of times that the corresponding binary production is drawn. $\zeta^{(k)} = \sum_{i,j} w^{(k)}_{i,j}$ approximates the log posterior well. When $w^{(k)}_{i,j}$ 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 $L_\theta - \eta \cdot \partial\mathcal{L}/\partial L_\theta$ where $\mathcal{L}$ is the log-posterior and η 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 ms overlap. 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: $\alpha = \gamma = 1$, $\rho = 1$, $\phi = \sqrt{2}$, $\sigma = 1$, $\eta = 0.1$, $\chi_1 = 3$, $\chi_2 = 2$, and $\lambda = \Omega T/\sum_{\omega,t} Y_{\omega,t}$. 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 and

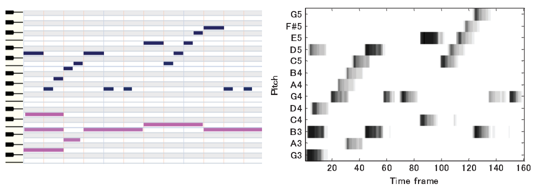

Fig. 2. Piano roll as the ground truth of note events (left) and estimated note events $\sum_n V_n O_{e_n, t-\tau_n}$ (right). We trained 13 spectral bases and 5 envelope patterns in advance. The proposed method captured the nearly adequate number of notes and their pitch information.

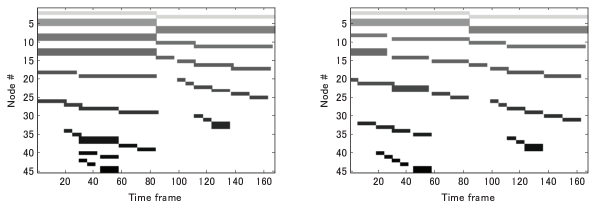

Fig. 3. Two examples of parsing trees corresponding to bars 2-4 of RWC-MDB-C-2001 No. 24A (shown in Fig. 2). Each row indicates the region (onset time and length) of each node of the estimated parsing tree on the time axis. Since 2-dimensional tree structures are too complicated to draw, parent-child 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.

$K=15$. The greatest length embedded in the symbols was fixed at the time length of segments. Fig. 2 shows that the proposed method decomposes audio signals into a nearly adequate number of active events with their pitch information. As shown in Fig. 3, the proposed sampler shows a meaningful hierarchical structure related to time-span associations and simultaneous events. Since the MCMC sampler gives us the estimated posterior distribution on parsing trees, the clue can be directly applied to automatic music transcription: for example the left sample implies that the opening four notes are simultaneous events in the form of an arpeggiated chord (played in rapid succession) in the music score.

For the second experiment we applied the proposed algorithm (truncation level: $K=20$) to two segments (bars 10-13 and bars 18-21) extracted manually from a jazz song consisting of piano, bass, and drums (RWC-MDB-J-2001 No. 16) [10]. It is very difficult to decompose these signals into individual note events. We thus set the sufficient number of events at $N=100$, and the parsing tree was estimated using the mainly effective 60 events that had a larger volume V_n. Because 40 smaller events contributed nothing to the parsing tree and their parameters, their onsets were estimated based only on Eq. (7) at the next iteration. Fig. 4 shows one example sampled by the proposed MCMC. The higher level captures the five frequent groups. "Swinging" rhythmics are captured in deeper layers. We confirmed that the proposed method could capture frequent patterns and simultaneous active events at several levels of hierarchy. Although instrumental information (pitch information) does not depend on tree structures in this paper, it must provide more meaningful representations of music signals. For example, we plan to learn frequent drum patterns and rhythmic structures of melodic voices simultaneously in the future.

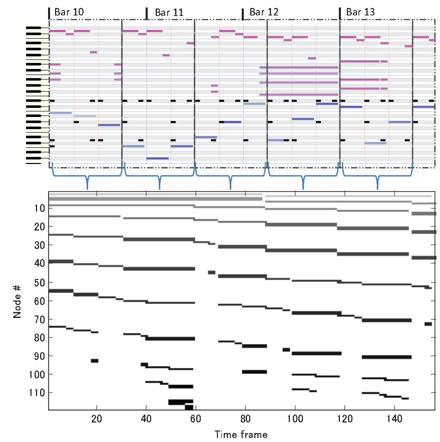

Fig. 4. Piano roll as the ground truth of note events (top) and estimated tree (bottom).

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 $\mathrm{Matlab}^{\mathrm{TM}}$. In the future, more sophisticated inference methods will be considered, such as collapsed sampling [8], slice sampler [11], and retrospective sampling [12].

7. REFERENCES

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

[2] S. Raczynski, N. Ono, and S. Sagayama, "Multipitch analisys with harmonic nonnegative matrix approximation," in *ISMIR*, 2007, pp. 381–386.

[3] F. Lerdahl and R. Jackendoff, "A Generative Theory of Tonal Music," *MIT Press*, 1983.

[4] M. Hamanaka and S. Tojo, "Interactive GTTM analyzer," in *ISMIR*, 2009, pp. 291–296.

[5] P. Liang, S. Petrov, M. I. Jordan, and D. Klein, "The infinite PCFG using hierarchical Dirichlet processes," in *EMNLP*, 2007, pp. 688–697.

[6] 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 *WASPAA*, 2011.

[7] J. Paisley, C. Wang, and D. Blei, "The Discreate infinite logistic Normal distribution for mixed-membership modeling," in *AISTATS*, 2011.

[8] M. Johnson, T. Griffiths, and S. Goldwater, "Bayesian inference for PCFGs via Markov chain Monte Carlo," in *NAACL*, 2007, pp. 139–146.

[9] O. Dikman and A. T. Cemgil, "Unsupervised single-channel source separation using Bayesian NMF," in *WASPAA*, 2009, pp. 93–96.

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

[11] M. Kalli, J. E. Griffin, and S. G. Walker, "Slice sampling mixture models," *Statistics and Computing*, vol. 21, no. 1, pp. 93–105, 2011.

[12] O. Papaspiliopoulos and G. O. Roberts, "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, pp. 169–186, 2008.