# UNSUPERVISED MUSIC UNDERSTANDING BASED ON NONPARAMETRIC BAYESIAN MODELS

*Kazuyoshi Yoshii* *Masataka Goto*

National Institute of Advanced Industrial Science and Technology (AIST)
{k.yoshii, m.goto}@aist.go.jp

## ABSTRACT

This paper presents a new research framework for *unsupervised music understanding*. Our goal is to recognize musical notes from polyphonic audio signals and simultaneously induce grammatical patterns from the recognized notes by integrating probabilistic acoustic and language models. Given music audio signals, both models could be jointly trained in a self-organizing manner without manually specifying the numbers of musical notes and grammatical patterns. In this paper, we introduce our nonparametric Bayesian acoustic and language models for multipitch analysis and chord progression analysis and discuss issues for integrating these models. We then provide a novel overview of various acoustic and language models whose underlying concepts are useful for implementing the framework.

***Index Terms***— Unsupervised music understanding, Bayesian nonparametrics, statistical machine learning, acoustic and language models, music transcription, grammar induction

## 1. INTRODUCTION

Music is one of the most sophisticated forms of audio signals. Even musically-untrained people can enjoy music and intuitively perceive that musical notes are organized to form sequential and simultaneous structures. Even if they have not been taught labels like C major and D minor, they intuitively know that only particular note combinations can sound harmonically (form chords). We assume that this musical sense can be acquired by just listening to a large amount of music, i.e., that people are capable of *unsupervised music understanding*. This capability is the basis of the analytical music listening that makes it easy for us to distinguish individual musical notes in familiar musical pieces. More specifically, humans intuitively grasp typical structural patterns that commonly appear in musical pieces they have listened to, and they use those patterns as grammatical clues for distinguishing overlapped notes. The recognized notes are in turn used in the deeper discovery of structural patterns.

In the light of the above discussion we propose a novel machine-learning framework for unsupervised music understanding (Fig. 1). We aim to recognize musical notes from polyphonic audio signals and simultaneously induce structural patterns from the recognized notes in an unsupervised manner. So far, converting continuous data (audio signals) into discrete data (musical notes) has been actively investigated, but pure audio-note modeling is insufficient as "music" signal processing because it does not consider grammatical patterns of musical notes (e.g. chords). We assume that what makes music "music" lies in the discrete world because most music can be denoted in symbolic representations (music scores). Inter-note modeling is thus indispensable for music understanding.

Our hierarchical framework is similar to the typical framework of automatic speech recognition (ASR), in which speech signals are transcribed into words (sequences of phonemes) by using an *acoustic model* that represents the spectral dynamics of speech signals and a *language model* that represents the connections between words. Such audio-phoneme modeling and inter-phoneme modeling are performed under the condition that all fixed vocabularies of linguistic units (e.g., phonemes and words) are explicitly given as an oracle. A key difference of "unsupervised" music understanding is that an adequate number of discrete "musical units" (e.g., notes and chords) should be inferred from continuous music audio signals without using any ground-truth vocabularies. This is analogous to language acquisition by infants, who are capable of discovering linguistic units in continuous speech signals in an unsupervised manner.

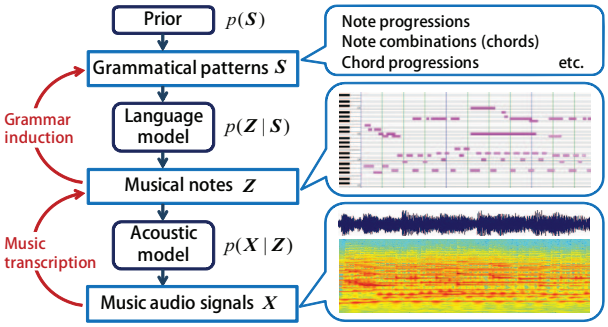


**Fig. 1**. A hierarchical Bayesian framework for unsupervised music understanding based on integration of acoustic and language models.

A principled approach to unsupervised music understanding is to formulate probabilistic acoustic and language models independently and then integrate them in a hierarchical Bayesian manner. A major difficulty in the recognition of musical notes is that the number of musical notes contained in audio signals is unknown. Therefore, the complexity of the acoustic model should be appropriately adjusted. In the induction of structural patterns, the complexity of the language model should also be adjusted according to the structural regularity of recognized notes. Conventional Bayesian models, however, force us to specify these complexities in advance even though the combinatorial search for optimal complexities is impractically expensive. *Nonparametric* Bayesian models, on the other hand, are free from this problem because they can not only learn their parameters (how likely each note or chord is to be used) but also efficiently adjust their own effective complexities (how many notes and chords should be considered) [1]. Such "musical units" can be generated unboundedly if they are needed to explain the given music data.

The remainder of this paper is organized as follows: Section 2 explains the probabilistic framework of unsupervised music understanding. Section 3 and Section 4 provide a novel overview of various acoustic and language models that could be used as components of the framework. Section 5 concludes the paper.

This study was supported in part by the JSPS KAKENHI 23700184.

## 2. UNSUPERVISED MUSIC UNDERSTANDING

We explain a mathematical formulation of the probabilistic framework for unsupervised music understanding. As shown in Fig. 1, the three kinds of random variables, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{S}$ respectively denote music audio signals or frequency spectra, musical notes, and structural patterns of those notes such as chords (note combinations) and progressions. We propose a fully hierarchical Bayesian model $p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S}) = p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z}|\boldsymbol{S})p(\boldsymbol{S})$, where $p(\boldsymbol{X}|\boldsymbol{Z})$ and $p(\boldsymbol{Z}|\boldsymbol{S})$ are called acoustic and language models, respectively, and $p(\boldsymbol{S})$ is a prior distribution over structural patterns. In this study $p(\boldsymbol{Z}|\boldsymbol{S})p(\boldsymbol{S})$ is regarded as a language model in a broad sense.

Our goal is to infer latent variables $\boldsymbol{Z}$ and $\boldsymbol{S}$ from observed data $\boldsymbol{X}$ in an unsupervised manner. For Bayesian inference, we aim to compute a posterior distribution over $\boldsymbol{Z}$ and $\boldsymbol{S}$ by using Bayes' rule, i.e., $p(\boldsymbol{S}, \boldsymbol{Z}|\boldsymbol{X}) = p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{S})/p(\boldsymbol{X})$, where a marginal likelihood (a.k.a. evidence) is given by $p(\boldsymbol{X}) = \int p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z}|\boldsymbol{S})p(\boldsymbol{S})d\boldsymbol{Z}d\boldsymbol{S}$. Although $p(\boldsymbol{X})$ is analytically intractable in general, we need to use approximate inference methods such as the variational Bayes (VB) and Markov-chain Monte Carlo (MCMC). If necessary, we can take the maximum-a-posteriori (MAP) point estimates of $\boldsymbol{Z}$ and $\boldsymbol{S}$ from the posterior distribution $p(\boldsymbol{S}, \boldsymbol{Z}|\boldsymbol{X})$. This data-driven framework enables us to compare different implementations of acoustic and language models in terms of the unified criterion $p(\boldsymbol{X})$. In addition, we can adapt this framework to a semi-supervised setting in which the size of a latent space is fixed and/or the values of latent variables are partially given as ground-truth definitions (musical knowledge).

One reason that we need nonparametric Bayesian models to implement $p(\boldsymbol{X}|\boldsymbol{Z})$ and $p(\boldsymbol{Z}|\boldsymbol{S})p(\boldsymbol{S})$ is that the space of $\boldsymbol{Z}$ and that of $\boldsymbol{S}$ are infinite in theory. Note that any additional knowledge (e.g., how many musical notes and chords are contained in observed data $\boldsymbol{X}$) is not available in the completely unsupervised setting. If an infinite amount of $\boldsymbol{X}$ were available, infinitely many kinds of musical notes and structural patterns would be needed to represent an infinite variety of $\boldsymbol{X}$. When the amount of $\boldsymbol{X}$ is finite, however, limited but unknown numbers of notes and patterns need to be considered A practically-important feature of nonparametric Bayesian models is that they theoretically have infinite complexity but actually instantiate only the numbers of musical notes and structural patterns necessary to represent $\boldsymbol{X}$. This enables us to manage such infinite models on real computers having finite computational power.

## 3. PROBABILISTIC ACOUSTIC MODELS

The goal of music transcription to convert polyphonic music audio signals into musical notes, i.e., discrete symbols that have absolute pitches (C0, C#0, D0, $\cdots$) and relative durations (whole, half, quarter, $\cdots$). Conventional acoustic models have been designed only for estimating fundamental frequencies (F0s) of musical sounds at every frame [2, 4, 5] or for separating musical sounds [6–11]. Some models [5] take into account the temporal continuity of frame-level F0s for detecting onsets and offsets of musical sounds. Note that the F0s and durations of musical sounds take continuous values represented by hertz and seconds. This means that it is implicitly assumed that at a post-processing stage the F0s are discretized by the semitone and the durations are quantized according to music tempo. This makes it difficult, however, to directly represent structural patterns (e.g., combinations and progressions) of discrete symbols (musical notes) in a probabilistic and principled manner.

In the framework of unsupervised music understanding, we need to build a new type of probabilistic acoustic models that can account for an infinite number of musical notes having infinitely many kinds of *discrete* pitches and durations. To represent $p(\boldsymbol{X}|\boldsymbol{Z})$, i.e., how audio spectra $\boldsymbol{X}$ are stochastically generated from musical notes $\boldsymbol{Z}$, several ideas of conventional acoustic models are worth considering. We explain two major types of acoustic models: *mixture models* and *factorial models*, which can be extended to nonparametric Bayesian models by taking the infinite limit of standard Bayesian models when the space of $\boldsymbol{Z}$ to diverges to infinity.

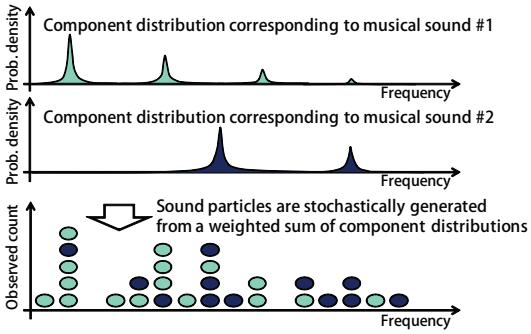


**Fig. 2**. Mixture modeling of amplitude spectrum.

### 3.1. Mixture Models

As shown in Fig. 2, the mixture modeling approach assumes that an amplitude spectrum is generated from *a weighted sum of probability distributions* that correspond to individual sounds. This means that an amplitude spectrum is interpreted as a histogram of "sound particles" having their own frequencies. If the amplitude value at frequency $f$ is $a$, we assume that a sound particle of frequency $f$ was observed $\lfloor a \rfloor$ times. Here, each particle is assumed to be generated from one of the component distributions. Although this assumption does not make physical sense, it is known to be useful in practice.

#### *3.1.1. Infinite Latent Harmonic Allocation*

One promising idea is to use a Gaussian mixture model (GMM) as a component distribution. More specifically, a harmonic structure consisting of $M$ harmonic partials is explicitly represented by binding component Gaussians to the frequencies of those partials. An amplitude spectrum consisting of $K$ harmonic sounds is represented as a mixture of GMMs. An underlying assumption is that each sound particle is stochastically generated from one of $KM$ Gaussians, i.e., one of $M$ partials contained in one of $K$ sounds. This idea is used by multi-F0 analyzers called PreFEst [4] and HTC [5], in which the values of $K$ and $M$ are assumed to be given in advance. The state-of-the-art analyzer called infinite latent harmonic allocation (iLHA) [2] is a nonparametric Bayesian version of PreFEst, in which the values of $K$ and $M$ are considered to be infinite in theory.

We explain a mathematical formulation of iLHA. Let $D$ be the number of frames and regard the observed spectra $\boldsymbol{X}$ as a set of frequencies observed over those frames, i.e., $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_D\}$. $\boldsymbol{X}_d = \{x_{d1}, \cdots, x_{dN_d}\}$ is a set of frequencies observed at frame $d$, where $N_d$ is the number of sound particles at that frame. Note that the value of $x_{dn}$ is represented on a logarithmic scale [cents]. Let $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_D\}$ and $\boldsymbol{Z}_d = \{\boldsymbol{z}_{d1}, \cdots, \boldsymbol{z}_{dN_d}\}$ be the corresponding latent variables, where $\boldsymbol{z}_{dn}$ is a $KM$-dimensional vector in which only one entry, $z_{dnkm}$, takes one and the others take zero when frequency $x_{dn}$ is generated from partial $m$ $(1 \leq m \leq M)$ of source $k$ $(1 \leq k \leq K)$. Therefore, the space of $\boldsymbol{Z}$ is defined over $K$ sounds and $M$ partials. Given $\boldsymbol{Z}$, the likelihood of $\boldsymbol{X}$ is defined as

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{dnkm} \mathcal{N}\left(x_{dn}\middle|\mu_k + 1200\log_2 m, \Lambda_k^{-1}\right)^{z_{dnkm}} \quad (1)$$

where $\mu_k$ is the F0 [cents] of harmonic sound $k$ and $\Lambda_k$ indicates a degree of sharpness of harmonic partials on the logarithmic scale. A

simple mixture model over $\boldsymbol{Z}$ is then formulated as follows:

$$p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{\tau}) = \prod_{dnkm} (\pi_{dk}\tau_{km})^{z_{dnkm}} \quad (2)$$

where $\pi_{dk}$ is a mixing ratio of sound $k$ at frame $d$ and $\tau_{km}$ is a mixing ratio of partial $m$ of sound $k$. Although this model is mathematically convenient for pure acoustic modeling, a musically-meaningful language model over $\boldsymbol{Z}$ should be developed in the future for unsupervised music understanding. To formulate a Bayesian model, we put conjugate priors over unknown parameters $\boldsymbol{\pi}$, $\boldsymbol{\tau}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$. More specifically, we use $K$-dimensional and $M$-dimensional Dirichlet priors as $p(\boldsymbol{\pi})p(\boldsymbol{\tau})$ and use Gauss-Gamma priors as $p(\boldsymbol{\mu},\boldsymbol{\Lambda})$. Given the observed data $\boldsymbol{X}$, the posterior distribution $p(\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}|\boldsymbol{X})$ can be computed by using a VB method.

To derive iLHA, we take the limit of $p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{\tau})p(\boldsymbol{\pi})p(\boldsymbol{\tau})$ as $K$ and $M$ approach infinity. $\boldsymbol{\pi}_d$ contains an infinite number of mixing ratios $\{\pi_{d1},\cdots,\pi_{d\infty}\}$ that still sum to unity. $p(\boldsymbol{\pi}_d)$ is an infinite-dimensional Dirichlet distribution, which is known to be equivalent to a Dirichlet process (DP). A recursive generative process of $\boldsymbol{\pi}_d$ is known as the stick-breaking process, where the values of $\boldsymbol{\pi}_d$ tend to decrease exponentially. Since in practice most values are too small, a limited number of harmonic sounds can appear in the observed data $\boldsymbol{X}$. A similar discussion can be applied to $\boldsymbol{\tau}$.

Alternatively, we can consider a marginal likelihood given by $p(\boldsymbol{Z})=\int(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{\tau})p(\boldsymbol{\pi})p(\boldsymbol{\tau})d\boldsymbol{\pi}d\boldsymbol{\tau}$, in which we do not need to directly deal with infinite-dimensional parameters $\boldsymbol{\pi}$ and $\boldsymbol{\tau}$. Although the space of $\boldsymbol{Z}$ is infinite in theory, limited kinds of harmonic sounds and partials appear in the finite data $\boldsymbol{X}$. A recursive generative process of $\boldsymbol{Z}$ is known as the Chinese restaurant process (CRP).

In our experiments using eight piano pieces of jazz and classical music [2], the frame-level accuracy of F0 estimation was 80.9%. The accuracy of PreFEst, on the other hand, was 78.5% even though the Dirichlet hyperparameters were carefully tuned. However, a limitation of iLHA is that the number of active sources (effective $K$) tends to be overestimated because the GMM is an oversimplified model of real harmonic sounds that contain not only harmonic partials but also noise components. This could be overcome by fusing iLHA with a language model to suppress unlikely note combinations.

#### *3.1.2. Infinite Probabilistic Latent Component Analysis*

Another promising idea is to use a two-dimensional discrete distribution on the time-frequency plane as a component distribution, which is factorized as the product of a discrete distribution over frequency bins and a discrete distribution over frames. These two distributions respectively correspond to the spectral shape and temporal activation of a sound source. This idea is known as probabilistic latent component analysis (PLCA) [6] and is useful for blind source separation. A nonparametric Bayesian variant [7] was recently proposed by letting the number of sound sources, $K$, diverge to infinity.

Here we explain a mathematical formulation of PLCA. Let $D$ and $F$ respectively be the numbers of frames and frequency bins. The observed spectra $\boldsymbol{X}$ can be regarded as a set of frame-bin pairs of sound particles contained in the spectra, i.e., $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$, where $N$ is the total number of particles. $\boldsymbol{x}_n$ is a $DF$-dimensional vector in which only one entry, $x_{ndf}$, takes one and the others take zero when the frame and frequency bin of the particle are equal to $d$ and $f$. Let $\boldsymbol{Z}=\{\boldsymbol{z}_1,\cdots,\boldsymbol{z}_N\}$ be the corresponding latent variables, where $z_n$ is a $K$-dimensional vector in which only one entry, $z_{nk}$, takes one and the others take zero when $\boldsymbol{x}_n$ is generated from source $k$ $(1\leq k\leq K)$. Therefore, the space of $\boldsymbol{Z}$ is defined over $K$ sound sources. Given $\boldsymbol{Z}$, the likelihood of $\boldsymbol{X}$ is defined as

$$p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\xi},\boldsymbol{\eta}) = \prod_{nkdf} (\xi_{kd}\eta_{kf})^{z_{nk}x_{ndf}} \quad (3)$$

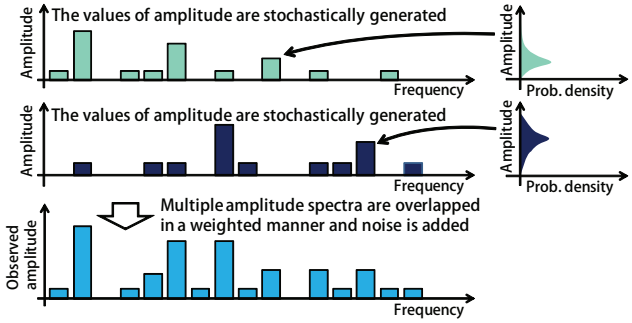


**Fig. 3**. Factorial modeling of amplitude spectrum.

where $\xi_{kd}$ is the probability that a sound particle of source $k$ is generated from frame $d$ and $\eta_{kf}$ is the probability that a sound particle of source $k$ is generated from frequency bin $f$. A simple mixture model over $\boldsymbol{Z}$ is then formulated as follows:

$$p(\boldsymbol{Z}|\boldsymbol{\omega}) = \prod_{nk} \omega_k^{z_{nk}} \quad (4)$$

where $\omega_k$ is a mixing ratio of sound $k$. Bayesian PLCA is easily formulated by using Dirichlet priors as $p(\boldsymbol{\xi})p(\boldsymbol{\eta})p(\boldsymbol{\omega})$. Nonparametric Bayesian treatment is also straightforward because $p(\boldsymbol{\omega})$ is equivalent to a DP when the value of $K$ approaches infinity.

### 3.2. Factorial models

As shown in Fig. 3, the factorial modeling approach assumes that an amplitude spectrum is approximated by *a weighted sum of multivariate random variables* that correspond to individual sounds. Nonnegative matrix factorization (NMF) [8] has been widely used among various factorial models such as principal component analysis (PCA) and independent component analysis (ICA). NMF regards a time-frequency spectrogram as a nonnegative matrix that can be factorized into the product of two nonnegative matrices, i.e., a set of spectral bases and a set of temporal activations.

Several nonparametric Bayesian models of NMF have been formulated by letting the number of spectral bases diverge to infinity. Hoffman *et al.* [9] proposed the GaP-NMF, which is an infinite extension of the Itakura-Saito divergence NMF [10]. Nakano *et al.* [11] formulated another GaP-NMF based on the Kullback-Leibler divergence NMF [8] and further extended it to allow each spectral basis to temporally vary according to an infinite hidden Markov model.

## 4. PROBABILISTIC LANGUAGE MODELS

The goal of grammar induction is to discover structural patterns $\boldsymbol{S}$ from musical notes $\boldsymbol{Z}$ for evaluating the structural appropriateness of those notes in terms of the likelihood $p(\boldsymbol{Z}|\boldsymbol{S})$. Conventionally, we are required to define as $\boldsymbol{S}$ a fixed and finite vocabulary of structural patterns such as chord types (what note combinations should be considered) and chord progressions (how many consecutive chords should be considered). However, appropriate definitions heavily depend on the structural complexity of target music data.

To solve this problem, we need to build nonparametric Bayesian models by using a prior $p(\boldsymbol{S})$ over an infinite number of structural patterns. We explain two types of language models: *chain-structured models* and *tree-structured models*, which can be extended to nonparametric Bayesian models when the space of $\boldsymbol{S}$ diverges to infinity.

### 4.1. Chain-structured Models

As shown in Fig. 4, the chain-structured modeling approach assumes that musical notes or chords vary according to Markovian dynamics. In standard chord progression analysis, $n$-gram models have been

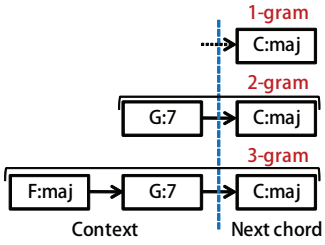


**Fig. 4**. Chain-structured modeling

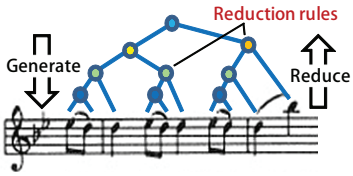


**Fig. 6**. Tree-structured modeling

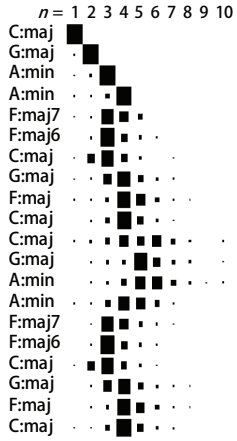


**Fig. 5**. Infinity-gram modeling for chord sequences

used for representing *sequences of conventional chord labels* (e.g., C major and D minor) by assuming each chord to depend on a context consisting of $n-1$ preceding chords [12]. This idea was taken from the field of computational linguistics (CL), in which $n$-gram models are often used for representing *sequences of words*. For example, Teh [13] proposed a hierarchical Pitman-Yor language model (HPYLM) as the first generative model of $n$-grams. Mochihachi and Sumita [14] proposed a variable-order Pitman-Yor language model (VPYLM or infinity-gram model) that allows each word (chord) to depend on an unbounded and variate number of preceding words, as shown in Fig. 5. However, the space of $\boldsymbol{S}$ is finite because a vocabulary of words (chord labels) are assumed to be specified.

We introduce a vocabulary-free infinity-gram model $p(\boldsymbol{Z}|\boldsymbol{S})$ [3] for *sequences of simultaneous musical notes* $\boldsymbol{Z}$. This model is an extended version of the VPYLM and does not force us to specify the value of $n$ and define a limited vocabulary of chord labels. The space of $\boldsymbol{S}$ is defined over infinitely many kinds of note combinations. A key building block is the Pitman-Yor process (PY), which is a prior distribution over distributions. Let $d$ and $\theta$ be positive scalars and $G_0$ be any distribution. The PY is represented as $G \sim \mathrm{PY}(d, \theta, G_0)$, where $d$ and $\theta$ are discount and strength parameters and $G$ is a random distribution. The larger the value of $\theta$ is, the more likely it is that $G$ is similar to $G_0$. Our $n$-gram model is obtained by layering PYs in a hierarchical Bayesian manner. Suppose we have an $n$-gram distribution $G_{\boldsymbol{u}}$ over $\boldsymbol{S}$, where $\boldsymbol{u}$ is a context of length $n-1$. An $n-1$-gram distribution $G_{\boldsymbol{u}^*}$ given the shortened context $\boldsymbol{u}^*$ is somewhat similar to $G_{\boldsymbol{u}}$. Here $G_{\boldsymbol{u}}$ is assumed to be drawn from a PY as $G_{\boldsymbol{u}} \sim \mathrm{PY}(d_n, \theta_n, G_{\boldsymbol{u}^*})$, where $d_n$ and $\theta_n$ are parameters specific to $n$. Such a process is defined recursively. Finally, the unigram distribution $G_\phi$ is given by $G_\phi \sim \mathrm{PY}(d_0, \theta_0, G_0)$, where $G_0$ is a global base measure (0-gram distribution) over $\boldsymbol{S}$. An important feature of our model is that $G_0$ itself is represented as a generative model $p(\boldsymbol{S})$ that evaluates how likely musical notes are to occur simultaneously.

In our comparative experiments using Beatles songs [3], the perplexity obtained with our model was 14.6, which was significantly better than that obtained with the VPYLM (15.8).

### 4.2. Tree-structured Models

As shown in Fig. 6, the tree-structured modeling approach assumes that a temporal sequence of musical notes can be hierarchically clustered according to reduction patterns $\boldsymbol{S}$. This concept is influenced by the Schenkerian theory. Hamanaka *et al.* [15] proposed a computational model for the generative theory of tonal music (GTTM), which is originated from the Schenkerian theory, where the conflicts between multiple reduction patterns are solved in an ad-hoc manner. Gilbert and Conklin [16] used a probabilistic context-free grammar (PCFG) for inferring a reduction tree from a melody line by regarding musical notes as words. Kirlin and Jensen [17] proposed a similar tree model for Schenkerian analysis. In the field of CL, Liang *et al.* [18] proposed a nonparametric Bayesian PCFG that can instantiate an unbounded number of reduction patterns as needed. This is highly promising for inducing a music grammar $p(\boldsymbol{S})$ itself.

## 5. CONCLUSION

This paper presented a hierarchical Bayesian framework for unsupervised music understanding. Our ultimate goal is to estimate musical notes from music audio signals and simultaneously induce structural patterns from the transcribed notes in an unsupervised way. In initial steps, we proposed state-of-the-art nonparametric Bayesian acoustic and language models. Future work includes fusing these models into a unified model. We consider exact Bayesian inference to be computationally tractable by alternately optimizing acoustic and language models with the Gibbs sampling method. In addition, to train a data-driven language model of music, our framework is expected to make use of a huge amount of music audio signals on the web even though those signals are *not* given any ground-truth transcriptions.

## 6. REFERENCES

[1] N. Hjort, C. Holmes, P. Müller, and S. Walker, Eds., *Bayesian Nonparametrics*, Cambridge University Press, 2010.

[2] K. Yoshii and M. Goto, "A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation," *IEEE Trans. on ASLP*, vol. 20, no. 3, pp. 717–730, 2012.

[3] K. Yoshii and M. Goto, "A vocabulary-free infinity-gram model for nonparametric Bayesian chord progression analysis," *ISMIR*, 2011.

[4] 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.

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

[6] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as non-negative factorizations," *Computational Intelligence and Neuoscience*, vol. 2008, 2008.

[7] M. Hoffman, D. Blei, and P. Cook, "Finding latent sources in recorded music with a shift-invariant HDP," *DAFX*, 2009.

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

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

[10] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[11] 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," *WASPAA*, 2011.

[12] R. Scholz, E. Vincent, and F. Bimbot, "Robust modeling of musical chord sequences using probabilistic N-grams," *ICASSP*, 2009, pp. 53–56.

[13] Y. W. Teh, "A Bayesian interpretation of interpolated Kneser-Ney," Tech. Rep. TRA2/06, National University of Singapore, 2006.

[14] D. Mochihashi and E. Sumita, "The infinite Markov model," *NIPS*, 2007, pp. 1017–1024.

[15] M. Hamanaka, K. Hirata, and S. Tojo, "FATTA: Full automatic time-span tree analyzer," *ICMC*, 2007, pp. 153–156.

[16] E. Gilbert and D. Conklin, "A probabilistic context-free grammar for melodic reduction," *IJCAI*, 2007, pp. 83–94.

[17] P. Kirlin and D. Jensen, "Probabilistic modeling of hierarchical music analysis," *ISMIR*, 2011, pp. 393–398.

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