A HIERARCHICAL BAYESIAN MODEL OF CHORDS, PITCHES, AND SPECTROGRAMS FOR MULTIPITCH ANALYSIS

Yuta Ojima[1] **Eita Nakamura**[1] **Katsutoshi Itoyama**[1] **Kazuyoshi Yoshii**[1]
[1] Graduate School of Informatics, Kyoto University , Japan
{ojima, enakamura}@sap.ist.i.kyoto-u.ac.jp, {itoyama, yoshii}@kuis.kyoto-u.ac.jp

ABSTRACT

This paper presents a statistical multipitch analyzer that can simultaneously estimate pitches and chords (typical pitch combinations) from music audio signals in an unsupervised manner. A popular approach to multipitch analysis is to perform nonnegative matrix factorization (NMF) for estimating the temporal activations of semitone-level pitches and then execute thresholding for making a piano-roll representation. The major problems of this cascading approach are that an optimal threshold is hard to determine for each musical piece and that musically inappropriate pitch combinations are allowed to appear. To solve these problems, we propose a probabilistic generative model that fuses an acoustic model (NMF) for a music spectrogram with a language model (hidden Markov model; HMM) for pitch locations in a hierarchical Bayesian manner. More specifically, binary variables indicating the existences of pitches are introduced into the framework of NMF. The latent grammatical structures of those variables are regulated by an HMM that encodes chord progressions and pitch co-occurrences (chord components). Given a music spectrogram, all the latent variables (pitches and chords) are estimated jointly by using Gibbs sampling. The experimental results showed the great potential of the proposed method for unified music transcription and grammar induction.

1. INTRODUCTION

The goal of automatic music transcription is to estimate the *pitches*, *onsets*, and *durations* of musical notes contained in polyphonic music audio signals. These estimated values must be directly linked with the elements of music scores. More specifically, in this paper, a pitch means a discrete fundamental frequency (F0) quantized in a semitone level, an onset means a discrete time point quantized on a regular grid (*e.g.*, eighth-note-level grid), and a duration means a discrete note value (integer multiple of the grid interval).

In this study we tackle multipitch estimation (subtask of automatic music transcription) that aims to make a binary piano-roll representation from a music audio signal, where only the existences of pitches are estimated at each frame. A popular approach to this task is to use non-negative matrix factorization (NMF) [1–7]. It approximates the magnitude spectrogram of an observed mixture signal as the product of a basis matrix (a set of basis spectra corresponding to different pitches) and an activation matrix (a set of temporal activations corresponding to those pitches). The existence of each pitch is then determined by executing thresholding or Viterbi decoding based a hidden Markov model (HMM) for the estimated activations [7, 8].

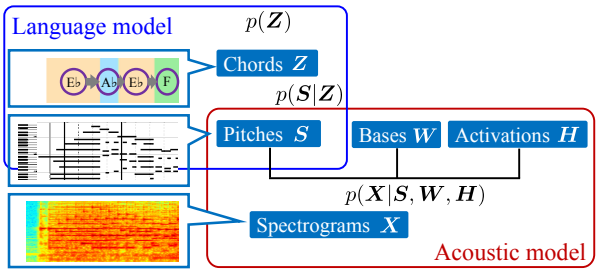

Figure 1. Overview of the proposed model consisting of language and acoustic models that are linked through binary variables $\boldsymbol{S}$ representing the existences of pitches.

This NMF-based cascading approach, however, has two major problems. First, it is hard to optimize a threshold for each musical piece. Second, the estimated results are allowed to be musically inappropriate because the relationships between different pitches are not taken into account. In fact, music has simultaneous and temporal structures; certain kinds of pitches (*e.g.*, C, G, and E) tend to simultaneously occur to form chords (*e.g.*, C major), which vary over time to form typical progressions. If such structural information is unavailable for multipitch analysis, we need to tackle the chicken-and-egg problem that chords are determined by pitch combinations, and vice versa.

To solve these problems, we propose a statistical method that can discover chords and pitches from music audio signals in an unsupervised manner while taking into account their interdependence (Fig.1). More specifically, we formulate a hierarchical Bayesian model that represents the generative process of an observed music spectrogram by unifying an *acoustic model* (probabilistic model underlying NMF) that represents how the spectrogram is generated from pitches and a *language model* (HMM) that represents how the pitches are generated from chords. A key feature of the unified model is that binary variables indicating the existences of pitches are introduced into the framework of NMF. This enables the HMM to represent both chord

transitions and pitch combinations using only discrete variables forming a piano-roll representation with chord labels. Given a music spectrogram, all the latent variables (pitches and chords) are estimated jointly by using Gibbs sampling.

The major contribution of this study is to realize unsupervised induction of musical grammars from music audio signals by unifying acoustic and language models. This approach is formally similar to, but essentially different from that to automatic speech recognition (ASR) because both the models are jointly learned in an unsupervised manner. In addition, our unified model has a three-level hierarchy (chord–pitch–spectrogram) while ASR is usually based on a two-level hierarchy (word–spectrogram). The additional layer is introduced by using an HMM instead of a Markov model (n-gram model) as a language model.

2. RELATED WORK

This section reviews related work on multipitch estimation (acoustic modeling) and on music theory implementation and musical grammar induction (language modeling).

2.1 Acoustic Modeling

The major approach to music signal analysi is to use nonnegative matrix factorization (NMF) [1–6, 9]. Cemgil *et al.* [9] developed a Bayesian inference scheme for NMF, which enabled the introduction of various hierarchical prior structures. Hoffman *et al.* [3] proposed a Bayesian nonparametric extension of NMF called gamma process NMF for estimating the number of bases. Liang *et al.* [6] proposed beta process NMF, in which binary variables are introduced to indicate the needs of individual bases at each frame. Another extension is source-filter NMF [4], which further decomposes the bases into sources (corresponding to pitches) and filters (corresponding to timbres).

2.2 Language Modeling

The implementation and estimation of music theory behind musical pieces are composed have been studied [10–12]. For example, some attempts have been made to computationally formulate the Generative Theory of Tonal Music (GTTM) [13], which represents the multiple aspects of music in a single framework. Hamanaka *et al.* [10] reformalized GTTM through a computational implementation and developed a method for automatically estimating a tree that represents the structure of music, called a timespan tree. Nakamura *et al.* [11] also re-formalized GTTM using a probabilistic context-free grammar model and proposed inference algorithms. These methods enabled automatic analysis of music. On the other hand, induction of music theory in an unsupervised manner has also been studied. Hu *et al.* [12] extended latent Dirichlet allocation and proposed a method for determining the key of a musical piece from symbolic and audio music based on the fact that the likelihood of appearance of each note tends to be similar among musical pieces in the same key. This method enabled the distribution of notes in a certain key to be obtained without using labeled training data.

Assuming that the concept of chords is a kind of music grammar, statistical methods of supervised chord recognition [14–17] are deeply related with unsupervised musical grammar induction. Rocher *et al.* [14] attempted chord recognition from symbolic music by constructing a directed graph of possible chords and then calculating the optimal path. Sheh *et al.* [15] used acoustic features called chroma vectors to estimate chords from music audio signals. They constructed an HMM whose latent variables are chord labels and whose observations are chroma vectors. Maruo *et al.* [16] proposed a method that uses NMF for extracting reliable chroma features. Since these methods need labeled training data, the concept of chords is required in advance. Approaches to make use of a sequence of chords in estimating pitches has also been proposed [18, 19]. This method estimates chord progressions and multiple pitches simultaneously by using a dynamic Bayesian network and shows better performance even with a simple acoustic model. Recent works employ recurrent neural networks as a language model to describe the relations between pitch combinations [20, 21].

3. PROPOSED METHOD

This section explains the proposed method of multipitch analysis that simultaneously estimates pitches and chords at the frame level from music audio signals. Our approach is to formulate a probabilistic generative model for observed music spectrograms and then solve the "inverse" problem, *i.e.*, given a music spectrogram, estimate unknown random variables involved in the model. The proposed model has a hierarchical structure consisting of acoustic and language models that are connected through a piano roll, *i.e.*, a set of binary variables indicating the existences of pitches (Fig. 1). The acoustic model represents the generative process of a music spectrogram from the piano roll, basis spectra, and temporal activations of individual pitches. The language model represents the generative process of chord progressions and pitch locations from chords.

3.1 Problem Specification

The goal of multipitch estimation is to make a piano roll from a music audio signal. Let $\boldsymbol{X} \in \mathbb{R}_{+}^{F \times T}$ be the magnitude spectrogram of a target signal, where F is the number of frequency bins and T is the number of time frames. We aim to convert $\boldsymbol{X}$ into a piano roll $\boldsymbol{S} \in \{0,1\}^{K \times T}$, which represents the existences of K kinds of pitches over T frames. In addition, we attempt to estimate a sequence of chords $\boldsymbol{Z} = \{z_t\}_{t=1}^{T}$.

3.2 Acoustic Modeling

The acoustic model is formulated in a similar way to beta-process NMF having binary masks [6] (Fig. 2). The given spectrogram $\boldsymbol{X} \in \mathbb{R}_{+}^{F \times T}$ is factorized into bases $\boldsymbol{W} \in \mathbb{R}_{+}^{F \times K}$, activations $\boldsymbol{H} \in \mathbb{R}_{+}^{K \times T}$, and binary variables $\boldsymbol{S} \in \{0,1\}^{K \times T}$ as follows:

$$X_{ft} | \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S} \sim \mathrm{Poisson}\left(\textstyle\sum_{k=1}^{K} W_{fk} H_{kt} S_{kt}\right), \quad (1)$$

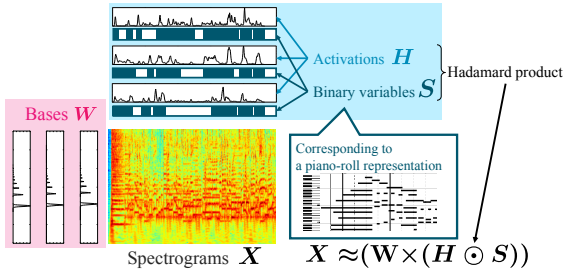

Figure 2. The overview of the acoustic model based on a variant of NMF having binary variables (masks).

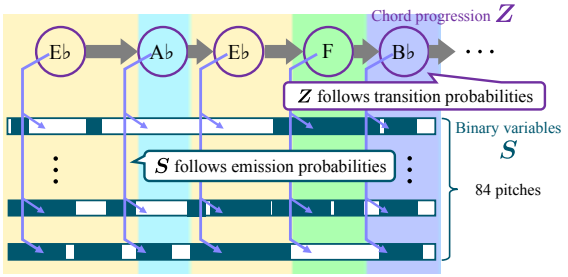

Figure 3. The overview of the language model based on an HMM that stochastically emits binary variables.

where $\{W_{fk}\}_{f=1}^{F}$ is the k-th basis spectrum, H_{kt} is the volume of basis k at frame t, and S_{kt} is a binary variable indicating whether or not basis k is used at frame t.

A set of basis spectra $\boldsymbol{W}$ is divided into two parts: harmonic spectra and noise spectra. In this study we prepare K_h harmonic basis spectra corresponding to K_h different pitches and one noise basis spectrum ($K = K_h + 1$). Assuming that the harmonic structures of the same instrument have the shift-invariant relationships, the harmonic part of $\boldsymbol{W}$ are given by

$$\{W_{fk}\}_{f=1}^{F} = \mathrm{shift}\left(\{W_f^{\mathrm{h}}\}_{f=1}^{F},\ \zeta(k-1)\right), \qquad (2)$$

for $k = 1, \ldots K_h$, where $\{W_f^{\mathrm{h}}\}_{f=1}^{F}$ is a harmonic template structure common to harmonic basis spectra used for NMF, $\mathrm{shift}\,(\boldsymbol{x}, a)$ is an operator that shifts $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ to $[0, \ldots, 0, x_1, \ldots, x_{n-a}]^T$, and ζ is the number of frequency bins corresponding to the semitone interval.

We put two kinds of priors on the harmonic template spectrum $\{W_f^{\mathrm{h}}\}_{f=1}^{F}$ and a noise basis spectrum $\{W_f^{\mathrm{n}}\}_{f=1}^{F}$. To make the harmonic spectrum sparse, we put a gamma prior on $\{W_f^{\mathrm{h}}\}_{f=1}^{F}$ as follows:

$$W_f^{\mathrm{h}} \sim \mathcal{G}\left(a^{\mathrm{h}}, b^{\mathrm{h}}\right) \qquad (3)$$

where a^{h} and b^{h} are hyperparameters. On the other hand, we put an inverse-gamma chain prior [22] on $\{W_f^{\mathrm{n}}\}_{f=1}^{F}$ to induce the spectral smoothness as follows:

$$G_f^W | W_{f-1}^{\mathrm{n}} \sim \mathcal{IG}\left(\eta^W, \tfrac{\eta^W}{W_{f-1}^{\mathrm{n}}}\right),$$
$$W_f^{\mathrm{n}} | G_f^W \sim \mathcal{IG}\left(\eta^W, \tfrac{\eta^W}{G_f^W}\right), \qquad (4)$$

where η^W is a hyperparameter that determines the strength of smoothness and G_f^W is an auxiliary variable that induces positive correlation between W_{f-1}^{n} and W_f^{n}.

A set of activations $\boldsymbol{H}$ is represented in the same way as $\boldsymbol{W}$. If H_{kt} takes almost zero, S_{kt} has no impact on NMF. This allows S_{kt} to take one (the corresponding pitch is judged to be activated) even though the activation H_{kt} is almost zero. We can avoid this problem by putting an inverse-gamma prior for H_{kt} to induce non-zero values. To induce the temporal smoothness in addition, we put the following inverse-gamma chain prior on $\boldsymbol{H}$:

$$G_{kt}^H | H_{k(t-1)} \sim \mathcal{IG}\left(\eta_H, \tfrac{\eta_H}{H_{k(t-1)}}\right),$$
$$H_{kt} | G_{kt}^H \sim \mathcal{IG}\left(\eta_H, \tfrac{\eta_H}{G_{kt}^H}\right), \qquad (5)$$

where η_H is a hyperparameter that determines the strength of smoothness and G_{kt}^H is an auxiliary variable that induces positive correlation between $H_{k(t-1)}$ and H_{kt}.

3.3 Language Modeling

The language model is an HMM that has a Markov chain of latent variables $\boldsymbol{Z} = \{z_1, \ldots, z_T\}$ ($z_t \in \{1, \ldots, I\}$) and emits binary variables $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T\}$ ($\boldsymbol{s}_t \in \{0,1\}^{K_h}$), where I represents the number of states (chords) and K_h represents the number of possible pitches. Note that $\boldsymbol{S}$ is actually a set of latent variables in the proposed unified model. The HMM is defined as:

$$z_1 | \boldsymbol{\phi} \sim \mathrm{Categorial}(\boldsymbol{\phi}), \qquad (6)$$
$$z_t | z_{t-1}, \boldsymbol{\psi}_{z_{t-1}} \sim \mathrm{Categorical}(\boldsymbol{\psi}_{z_{t-1}}), \qquad (7)$$
$$S_{kt} | z_t, \pi_{z_t k} \sim \mathrm{Bernoulli}(\pi_{z_t k}) \qquad (8)$$

where $\boldsymbol{\psi}_i \in \mathbb{R}^I$ is a set of transition probabilities of chord i, $\boldsymbol{\phi} \in \mathbb{R}^I$ is a set of initial probabilities, and $\pi_{z_t k}$ indicates the probability that the k-th pitch is emitted under a chord z_t, We put conjugate priors on these parameters as:

$$\boldsymbol{\psi}_i \sim \mathrm{Dir}(\mathbf{1}_I), \quad \boldsymbol{\phi} \sim \mathrm{Dir}(\mathbf{1}_I), \quad \pi_{z_t k} \sim \mathrm{Beta}(e, f), \qquad (9)$$

where $\mathbf{1}_I$ is the I-dimensional all-one vector and e and f are hyperparameters.

In practice, we represent only the emission probabilities of 12 pitch classes (C, C#, . . ., B) in one octave. Those probabilities are copied and pasted to recover the emission probabilities of K_h kinds of pitches. In addition, the emission probabilities $\{\pi_{ik}\}_{k=1}^{K_h}$ of chord i are forced to have circular-shifting relationships with those of other chords of the same type. In this paper, we consider only major and minor chords as chord types ($I = 2 \times 12$) for simplicity.

3.4 Posterior Inference

Given the observed data $\boldsymbol{X}$, our goal is to calculate the posterior distribution $p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\psi} | \boldsymbol{X})$. Since analytic calculation is intractable, we use Markov chain Monte Carlo (MCMC) methods as in [23]. Since the acoustic and language models share only the binary variables, each model can be updated independently when the binary variables are given. These models and binary variables are iteratively sampled. Finally, the latent variables (chord progressions) of the language model are estimated by using the Viterbi algorithm and the binary variables (pitch locations) are determined by using parameters having the maximum likelihood.

3.4.1 Sampling Binary Variables

The binary variables $\boldsymbol{S}$ are sampled from a posterior distribution that is calculated by integrating the acoustic model

as a likelihood function and the language model as a prior distribution according the Bayes' rule. Note that as shown in Fig. 1, the binary variables $\boldsymbol{S}$ are involved in both acoustic and language models (*i.e.*, the probability of each pitch being used is determined by a chord, and whether or not each pitch is used affects the reconstructed spectrogram). The conditional posterior distribution of S_{kt} is given by

$$S_{kt} \sim \text{Bernoulli}\left(\frac{P_1}{P_1+P_0}\right), \tag{10}$$

where P_1 and P_0 are given by

$$\begin{aligned} P_1 &= p(S_{kt}=1|S_{\neg k,t}, \boldsymbol{x}_t, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\pi}, \boldsymbol{z}, \alpha) \qquad (11)\\ &\propto \pi_{z_k}^{\alpha}\prod_f \left(\hat{X}_{ft}^{\neg k} + W_{fk}H_{kt}\right)^{X_{ft}} \exp\{-W_{fk}H_{kt}\},\\ P_0 &= p(S_{kt}=0|S_{\neg k,t}, \boldsymbol{x}_t, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\pi}, \alpha)\\ &\propto (1-\pi_{z_k})^{\alpha}\prod_f \left(\hat{X}_{ft}^{\neg k}\right)^{X_{ft}}, \qquad (12) \end{aligned}$$

where $\hat{X}_{ft}^{\neg k} \equiv \sum_{l\neq k} W_{fl}H_{lt}S_{lt}$ denotes the magnitude at frame t reconstructed without using the k-th basis and α is a parameter that determines the weight of the language model relative to that of the acoustic model. Such a weighting factor is also needed in ASR. If α is not equal to one, Gibbs sampling cannot be used because the normalization factor cannot be analytically calculated. Instead, the Metropolis-Hastings (MH) algorithm is used by regarding Eq. (10) is used as a proposal distribution

3.4.2 Updating the Acoustic Model

The parameters of the acoustic model $\boldsymbol{W}^{\mathrm{h}}$, $\boldsymbol{W}^{\mathrm{n}}$, and $\boldsymbol{H}$ can be sampled using Gibbs sampling. These parameters are categorized into those having gamma priors ($\boldsymbol{W}^{\mathrm{h}}$) and those inverse-gamma chain priors ($\boldsymbol{W}^{\mathrm{n}}$ and $\boldsymbol{H}$).

Using the Bayes' rule, the conditional posterior distribution of $\boldsymbol{W}^{\mathrm{h}}$ is given by

$$W_{fk}^{\mathrm{h}} \sim \mathcal{G}\left(\sum_t X_{ft}\lambda_{ftk} + a^{\mathrm{h}},\ \sum_t H_{kt}S_{kt} + b^{\mathrm{h}}\right), \tag{13}$$

where λ_{ftk} is a normalized auxiliary variable that is calculated with the latest sampled variables $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{H}}$, and $\hat{\boldsymbol{S}}$, as:

$$\lambda_{ftk} = \frac{\hat{W}_{fk}\hat{H}_{kt}\hat{S_{kt}}}{\sum_l \hat{W}_{fl}\hat{H}_{lt}\hat{S_{lt}}}. \tag{14}$$

The other parameters are sampled through auxiliary variables. Since $\boldsymbol{H}$ and $\boldsymbol{G}^H$ are interdependent in Eq. (5) and cannot be sampled jointly, $\boldsymbol{G}^H$ and $\boldsymbol{H}$ are sampled altenately. The conditional posterior of $\boldsymbol{G}^H$ is given by

$$G_{kt}^H \sim \mathcal{IG}\left(2\eta_H, \eta_H\left(\frac{1}{H_{kt}} + \frac{1}{H_{k(t-1)}}\right)\right). \tag{15}$$

Similarly, the conditional posteriors of $\boldsymbol{H}$, $\boldsymbol{G}^W$, and $\boldsymbol{W}^{\mathrm{n}}$ are given by

$$H_{kt} \sim \mathcal{IG}\left(2\eta_H, \eta_H\left(\frac{1}{G_{k(t+1)}^H} + \frac{1}{G_{kt}^H}\right)\right), \tag{16}$$

$$G_f^W \sim \mathcal{IG}\left(2\eta_W, \eta_W\left(\frac{1}{W_f^{\mathrm{n}}} + \frac{1}{W_{f-1}^{\mathrm{n}}}\right)\right), \tag{17}$$

$$W_f^{\mathrm{n}} \sim \mathcal{IG}\left(2\eta_W, \eta_W\left(\frac{1}{G_{f+1}^W} + \frac{1}{G_f^W}\right)\right), \tag{18}$$

if the observation $\boldsymbol{X}$ is not taken into account. Using the Bayes' rule and Jensen's inequality as in Eq. (13) and regarding Eq. (16) as a prior, the conditional posterior considering the observation $\boldsymbol{X}$ is written as follows: [1]

$$H_{kt} \sim \text{GIG}\left(2S_{kt}\textstyle\sum_f W_{fk}, \delta_H, \sum_f X_{ft}\lambda_{ftk} - \gamma_H\right),$$

where $\gamma_H = 2\eta_H$ and $\delta_H = \eta_H(\frac{1}{G_{k(t+1)}^H} + \frac{1}{G_{kt}^H})$. The conditional posterior of $\boldsymbol{W}^{\mathrm{n}}$ can be derived in the same manner as follows:

$$W_{fk}^{\mathrm{n}} \sim \text{GIG}\left(2\textstyle\sum_t H_{kt}S_{kt}, \delta_W, \sum_t X_{ft}\lambda_{ftk} - \gamma_W\right),$$

where $\gamma_W = 2\eta_W$ and $\delta_W = \eta_W(\frac{1}{G_{f+1}^W} + \frac{1}{G_f^W})$

3.4.3 Updating the Language Model

The latent variables $\boldsymbol{Z}$ are sampled from the following conditional posterior distribution:

$$p(z_t|\boldsymbol{S}, \boldsymbol{\pi}, \boldsymbol{\phi}, \Psi) \propto p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_t, z_t), \tag{19}$$

where $\boldsymbol{\pi}$ is the emission probabilities, $\boldsymbol{\phi}$ is the initial probabilities, and $\Psi = \{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_I\}$ is a set of the transition probabilities from each state. The right-hand side of Eq. (19) is further factorized using the conditional independence over $\boldsymbol{Z}$ and $\boldsymbol{S}$ as follows:

$$\begin{aligned} &p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_t, z_t)\\ &= p(\boldsymbol{s}_t|z_t)\textstyle\sum_{z_{t-1}} p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{t-1}, z_{t-1})p(z_t|z_{t-1}), \qquad (20)\\ &p(\boldsymbol{s}_1, z_1) = p(z_1)p(\boldsymbol{s}_1|z_1) = \phi_{z_1}p(\boldsymbol{s}_1|\pi_{z_1}). \qquad (21) \end{aligned}$$

Using Eqs. (20) and (21) recursively, $p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T|z_T)$ can be efficiently calculated via forward filtering and the last variable z_T is sampled according to $z_T \sim p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T|z_T)$. If the latent variables $z_{t+1}, \ldots, z_T$ are given, z_t is sampled from a posterior given by

$$p(z_t|\boldsymbol{S}, z_{t+1}, \ldots, z_T) \propto p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_t, z_t)p(z_{t+1}|z_t). \tag{22}$$

Since $p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_t, z_t)$ can be calculated in Eq. (20), z_t is recursively sampled from $z_t \sim p(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_t, z_t)p(z_{t+1}|z_t)$ via backward sampling.

The posterior distribution of the emission probabilities $\boldsymbol{\pi}$ is given by using the Bayes' rule as follows:

$$p\left(\boldsymbol{\pi}|\boldsymbol{S}, \boldsymbol{z}, \boldsymbol{\phi}, \Psi\right) \propto p\left(\boldsymbol{S}|\boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\phi}, \Psi\right) p\left(\boldsymbol{\pi}\right). \tag{23}$$

This is analytically calculable because $p\left(\boldsymbol{\pi}\right)$ is a conjugate prior of $p\left(\boldsymbol{S}|\boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\phi}, \Psi\right)$. Let C_i be the number of occurrences of chord $i \in \{1 \ldots I\}$ in $\boldsymbol{Z}$ and $\boldsymbol{c}_i \equiv \sum_{t\in\{t|z_t=i\}} \boldsymbol{s}_t$ be a K-dimensional vector that denotes the sum of $\boldsymbol{s}_t$ under the condition $z_t = i$. The parameters $\boldsymbol{\pi}$ are sampled according to a conditional posterior given by

$$\boldsymbol{\pi} \sim \text{Beta}\left(e + c_{ik}, f + C_i - c_{ik}\right). \tag{24}$$

The posterior distributions of the transition probabilities $\boldsymbol{\psi}$ and the initial probabilities $\boldsymbol{\pi}$ are given similarly as follows:

$$\begin{aligned} &p(\boldsymbol{\phi}|\boldsymbol{S}, \boldsymbol{z}, \boldsymbol{\pi}, \Psi) \propto p(z_1|\boldsymbol{\phi})\, p(\boldsymbol{\phi}) \qquad (25)\\ &p(\boldsymbol{\psi}|\boldsymbol{S}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\phi}) \propto \textstyle\prod_t p(z_t|z_{t-1}, \boldsymbol{\psi}_{z_{t-1}})\, p(\boldsymbol{\psi}_{z_{t-1}}). \qquad (26) \end{aligned}$$

Since $p(\boldsymbol{\phi})$ and $p\left(\boldsymbol{\psi}_i\right)$ are conjugate priors of $p(z_1|\boldsymbol{\phi})$ and $p(z_t|z_{t-1}, \boldsymbol{\psi}_{z_{t-1}})$, respectively, these posteriors can be easily calculated. Let $\boldsymbol{e}_i$ be the unit vector whose i-th element

[1] $\text{GIG}(a, b, p) \equiv \frac{(a/b)^{\frac{p}{2}}}{2K_p(\sqrt{ab})} x^{p-1} \exp(-\frac{ax+\frac{b}{x}}{2})$ denotes a generalized inverse Gaussian distribution.

is 1 and $\boldsymbol{a}_i$ be the I-dimensional vector whose j-th element denotes the number of transition from state i to state j. The parameters $\boldsymbol{\phi}$ and $\boldsymbol{\psi}_i$ are sampled according to conditional posteriors given by

$$\boldsymbol{\phi} \sim \mathrm{Dir}\left(\mathbf{1}_I + \boldsymbol{e}_{z_1}\right), \quad \boldsymbol{\psi}_i \sim \mathrm{Dir}\left(\mathbf{1}_I + \boldsymbol{a}_i\right). \tag{27}$$

4. EVALUATION

We report comparative experiments we conducted to evaluate the performance of our proposal model in pitch estimation. First, we confirmed in a preliminary experiment that correct chord progressions and emission probabilities were estimated from the piano-roll by the language model. Then, we estimated the piano-roll representation from acoustic audio signals by using the hierarchical model and the acoustic model.

4.1 Experimental Conditions

We used 30 pieces (labeled as "ENSTDkCl") selected from the MAPS database [24]. We converted them into monaural signals and truncated each of them to 30 seconds from the beginning. The magnitude spectrogram was made by using the variable-Q transform [25]. The 926×10075 spectrogram thus obtained was resampled to 926×3000 by using MATLAB's resample function. Moreover, we used harmonic and percussive source separation (HPSS) [26] as a preprocessing. Unlike the original study, HPSS was performed in the log-frequency domain. Median filter is applied over 50 time frames and 40 frequency bins each. Hyperparameters were empirically determined as $I = 24, a^{\mathrm{h}} = 1, b^{\mathrm{h}} = 1, a^{\mathrm{n}} = 2, b^{\mathrm{n}} = 1, c = 2, d = 1, e = 5, f = 80, \alpha = 1300, \eta_W = 800000$ and $\eta_H = 15000$. The emission probabilities are obtained for 12 notes, which are expanded to cover 84 pitches. In practice, we fixed the probability of internal transition (*i.e.* $p(z_{t+1} = z_t | z_t)$) to a large value ($1 - 8.0 \times 10^{-8}$) and assumed that the probabilities of transition to a different state follow Dirichlet distribution as shown in section 3.4.3 We implemented the proposed method by using C++ and a linear algebra library called Eigen3. The estimation was conducted with a standard desktop computer with an Intel Core i7-4770 CPU (8-core, 3.4 GHz) and 8.0 GB of memory. The processing time for the proposed method with one music piece (30 seconds as mentioned above) was 15.5 minutes .

4.2 Chord Estimation for Piano Rolls

We first verified that the language model properly estimated the emission probabilities and a chord progression. As an input, we combined correct binary piano-roll representations for 84 pitches (MIDI numbers 21–104) of the pieces we used. Since each representation has 3000 time-frames and we used 30 pieces, the input was 84×90000 matrix. We evaluated the precision of chord estimation as the ratio of the number of frames whose chords were estimated correctly to the total number of frames. Since we prepared two chord types for each root note, we treated "major" and "7th" in the ground-truth chords as "major" in the estimated chords, and "minor" and "minor 7th" in the ground-truth chords as "minor" in the estimated chords. In evaluation, other chord types were not used in evaluation and chord labels were estimated to maximize the precision since we estimated chords in an unsupervised manner. Since original MAPS database doesn't contain chord information, one of the authors labeled chord information for each music piece by hand [2].

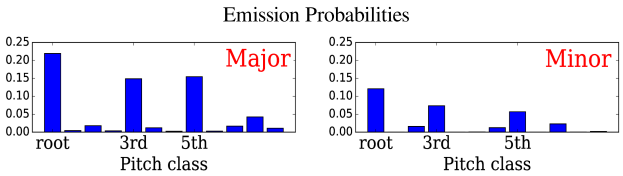

Figure 4. Emission probabilities estimated in the preliminary experiment. The left corresponds to major chords and the right corresponds to minor chords.

The experimental results shown in Fig. 4 shows that major chords and minor chords, which are typical chord types in tonal music, were obtained as emission probabilities. This implies that we can obtain the concept of chord from piano-roll data without any prior knowledge. The precision was 61.33%, which indicates our model estimates chords correctly to some extent even in an unsupervised manner. On the other hand, other studies on chord estimation have reported higher score [15, 16]. This is because that they used labeled training data and that they evaluated their method with popular music, which has clearer chord structure than classical music we used.

4.3 Multipitch Estimation for Music Audio Signals

We then evaluated our model in terms of the frame-level recall/precision rates and F-measure:

$$\mathcal{R} = \frac{\sum_t c_t}{\sum_t r_t}, \ \mathcal{P} = \frac{\sum_t c_t}{\sum_t e_t}, \ \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}}, \tag{28}$$

where r_t, e_t, and c_t are respectively the numbers of ground truth, estimated and correct pitches at the t-th time-frame. To cope with the arbitrariness in octaves of the obtained bases, estimated results for the whole piece were shifted by octaves and the most accurate one was used for the evaluation. We conducted a few comparative experiments under the following conditions: 1) Chords were fixed and unchanged during a piece (the acoustic model), 2) the language model was pre-trained using the correct chord labels and a correct piano-roll, and the learned emission probabilities were used in estimation (pre-trained with chord), 3) the language model was pre-trained using only a correct piano-roll, and the learned emission probabilities were used in estimation (pre-trained without chord). we evaluated the performances under the second and the third conditions by using cross-validation.

As shown in Table 1, the performance of the proposed method in the unsupervised setting (65.0%) was better than that of the acoustic model (64.7%). As shown in Fig. 5, the F-measure improvement due to integrating the language model for each piece correlated positively with the preci-

[2] The annotation data used for evaluation is available on http://sap.ist.i.kyoto-u.ac.jp/members/ojima/mapschord.zip

Condition	$\mathcal{F}$	$\mathcal{R}$	$\mathcal{P}$
The integrated model	65.0	67.3	62.8
The acoustic model	64.7	64.7	64.7
Pre-trained w/ chord	65.5	65.3	65.6
Pre-trained w/o chord	65.0	65.5	64.6

Table 1. Experimental results of multipitch analysis for 30 piano pieces labeled as ENSTDkCl.

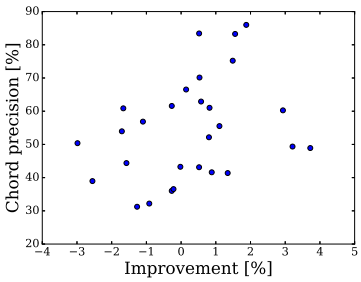

Figure 5. Correlation between estimated chord precision and the improvement of F-measure.

sion of chord estimation for each piece (correlation coefficient $r = 0.33$). This indicates that refining the language model also improves the pitch estimation.

Moreover, as shown in Fig. 6, major and minor chords like those in Fig. 4 were obtained as emission probabilities directly from music audio signals without any prior knowledge. This implies that frequently used chord types can be inferred from music audio signals automatically, which would be useful in music classification or similarity analysis. The performance in the supervised setting (65.5%) was better than the performance obtained in the unsupervised settings. Since there exist published piano scores with chord labels, this setting is considered to be practical. Although this difference was statistically insignificant (standard error was about 1.5%), F-measures were improved for 25 pieces out of 30. Moreover, the improvement exceeded 1% for 15 pieces. The example of pitch estimation shown in Fig. 7 indicates that insertion errors at low pitches are reduced by integrating the language model. On the other hand, insertion errors in total increased in the integrated model. This is because the constraint on harmonic partials (shift-invariant) is too strong to appropriately estimate the spectrum of each pitch. As a result, the overtones that should be expressed by a single pitch are expressed by multiple inappropriate pitches that do not exist in the ground-truth.

There would be much room for improving the performance. The acoustic model has the strong constraint on harmonic partials as mentioned above. This constraint can be relaxed by introducing source-filter NMF [4], which further decomposes the bases into sources corresponding to pitches and filters corresponding to timbres. Our model corresponds the case the number of filters is one, and increment of the number of filters would contribute to express difference in timbres (*e.g.*, difference between the timbre of high pitches and that of low pitches). The language model, on the other hand, can be refined by introducing other music theory such as keys. Some methods that treat the relationship between keys and chords [27], or keys and notes [12], have been studied. Moreover, the language model focus on reducing unmusical errors such as insertion errors in adjacent pitches, and is difficult to cope with errors in octaves or overtones. Modeling transitions between notes (horizontal relations) will contribute to solve this problem and to improve the accuracy.

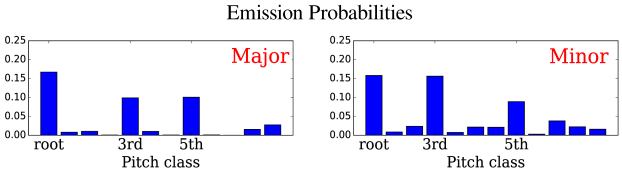

Figure 6. Emission probabilities learned from estimated piano-roll. Chord structures like those in Fig. 4 were obtained.

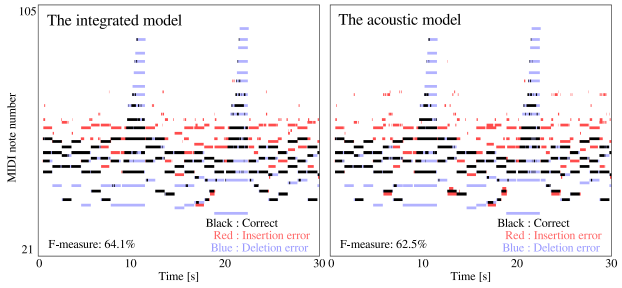

Figure 7. Estimated piano-rolls for MUS-bk_xmas5_ENSTDkCl. Integrating the language model redeuced Insertion errors at low pitches.

5. CONCLUSION

We presented a new statistical multipitch analyzer that can simultaneously estimate pitches and chords from music audio signals. The proposed model consists of an acoustic model (a variant of Bayesian NMF) and a language model (Bayesian HMM), and each model can make use of each other's information. The experimental results showed the potential of the proposed method for unified music transcription and grammar induction from music audio signals. On the other hand, each model has much room for performance improvement: the acoustic model has a strong constraint, and the language model is insufficient to express music theory. Therefore, we plan to introduce a source-filter model as the acoustic model and to introduce the concept of key in the language model.

Our approach has a deep connection to language acquisition. In the field of natural language processing (NLP), unsupervised grammar induction from a sequence of words and unsupervised word segmentation for a sequence of characters have actively been studied [28,29]. Since our model can directly infer musical grammars (*e.g.*, concept of chords) from either music scores (discrete symbols) or music audio signals, the proposed technique is expected to be useful for an emerging topic of language acquisition from continuous speech signals [30].

Acknowledgement: This study was partially supported by JST OngaCREST Project, JSPS KAKENHI 24220006, 26700020, 26280089, 16H01744, and 15K16054, and Kayamori Foundation.''

6. REFERENCES

[1] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE WASPAA*, pages 177–180, 2003.

[2] K. Ohanlon, H. Nagano, N. Keriven, and M. Plumbley. An iterative thresholding approach to L0 sparse hellinger NMF. In *ICASSP*, pages 4737–4741, 2016.

[3] M. Hoffman, D. M. Blei, and P. R. Cook. Bayesian nonparametric matrix factorization for recorded music. In *ICML*, pages 439–446, 2010.

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

[5] J. L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE TASLP*, 18(3):564–575, 2010.

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

[7] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE TASLP*, 18(3):528–537, 2010.

[8] G. E. Poliner and D. P. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Applied Signal Processing*, 2007.

[9] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

[10] M. Hamanaka, K. Hirata, and S. Tojo. Implementing a generative theory of tonal music. *Journal of New Music Research*, 35(4):249–277, 2006.

[11] E. Nakamura, M. Hamanaka, K. Hirata, and K. Yoshii. Tree-structured probabilistic model of monophonic written music based on the generative theory of tonal music. In *ICASSP*, 2016.

[12] D. Hu and L. K. Saul. A probabilistic topic model for unsupervised learning of musical key-profiles. In *ISMIR*, pages 441–446, 2009.

[13] R. Jackendoff and F. Lerdahl. *A generative theory of tonal music*. MIT Press, 1985.

[14] M. Rocher, T.and Robine, P. Hanna, and R. Strandh. *Dynamic Chord Analysis for Symbolic Music*. Ann Arbor, MI: MPublishing, University of Michigan Library, 2009.

[15] A. Sheh and D. P. Ellis. Chord segmentation and recognition using EM-trained hidden Markov models. In *ISMIR*, pages 185–191, 2003.

[16] S. Maruo, K. Yoshii, K. Itoyama, M. Mauch, and M. Goto. A feedback framework for improved chord recognition based on NMF-based approximate note transcription. In *ICASSP*, pages 196–200, 2015.

[17] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *ICASSP*, pages 5518–5521, 2010.

[18] S. Raczynski, E. Vincent, F. Bimbot, and S. Sagayama. Multiple pitch transcription using DBN-based musicological models. In *ISMIR*, pages 363–368, 2010.

[19] S. A. Raczynski, E. Vincent, and S. Sagayama. Dynamic bayesian networks for symbolic polyphonic pitch modeling. *IEEE TASLP*, 21(9):1830–1840, 2013.

[20] S. Sigtia, E. Benetos, and S. Dixon. An end-to-end neural network for polyphonic piano music transcription. *IEEE TASLP*, 24(5):927–939, 2016.

[21] S. Sigtia, E. Benetos, S. Cherla, T. Weyde, A. Garcez, and S. Dixon. An RNN-based music language model for improving automatic music transcription. In *ISMIR*, pages 53–58, 2014.

[22] A. T. Cemgil and O. Dikmen. Conjugate Gamma Markov random fields for modelling nonstationary sources. In *Independent Component Analysis and Signal Separation*, pages 697–705. Springer, 2007.

[23] M. Davy and S. J. Godsill. Bayesian harmonic models for musical signal analysis. *Bayesian Statistics*, (7):105–124, 2003.

[24] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE TASLP*, 18(6):1643–1654, 2010.

[25] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler. A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution. In *Audio Engineering Society Conference*, 2014.

[26] D. Fitzgerald. Harmonic/percussive separation using median filtering. In *DAFx*, pages 1–4, 2010.

[27] K. Lee and M. Slaney. Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio. *IEEE TASLP*, 16(2):291–301, 2008.

[28] M. Johnson. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, pages 398–406, 2008.

[29] D. Mochihashi, T. Yamada, and N. Ueda. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *ACL*, pages 100–108. Association for Computational Linguistics, 2009.

[30] T. Taniguchi and S. Nagasaka. Double articulation analyzer for unsegmented human motion using Pitman-Yor language model and infinite hidden markov model. In *IEEE/SICE International Symposium on System Integration*, pages 250–255. IEEE, 2011.