SIP (2018), vol. 7, e14, page 1 of 14 © The Authors, 2018.

doi:10.1017/ATSIP.2018.17

ORIGINAL PAPER

Chord-aware automatic music transcription based on hierarchical Bayesian integration of acoustic and language models

YUTA OJIMA, EITA NAKAMURA, KATSUTOSHI ITOYAMA AND KAZUYOSHI YOSHII

This paper describes automatic music transcription with chord estimation for music audio signals. We focus on the fact that concurrent structures of musical notes such as chords form the basis of harmony and are considered for music composition. Since chords and musical notes are deeply linked with each other, we propose joint pitch and chord estimation based on a Bayesian hierarchical model that consists of an acoustic model representing the generative process of a spectrogram and a language model representing the generative process of a piano roll. The acoustic model is formulated as a variant of non-negative matrix factorization that has binary variables indicating a piano roll. The language model is formulated as a hidden Markov model that has chord labels as the latent variables and emits a piano roll. The sequential dependency of a piano roll can be represented in the language model. Both models are integrated through a piano roll in a hierarchical Bayesian manner. All the latent variables and parameters are estimated using Gibbs sampling. The experimental results showed the great potential of the proposed method for unified music transcription and grammar induction.

Keywords: Automatic Music Transcription, Chord Estimation, Non-negative Matrix Factorization, Bayesian Inference

Received 13 July 2018; Revised 10 October 2018

I. INTRODUCTION

Automatic music transcription (AMT) refers to the estimation of pitches, onset times, and durations of musical notes from music signals and has been considered to be important for music information retrieval. Since multiple pitches usually overlap in polyphonic music and each pitch consists of many overtone components, estimation of multiple pitches is still an open problem. Although such multipitch estimation is often called AMT, quantization of onset times and durations of musical notes is required for completing AMT.

A major approach to multipitch estimation and AMT is to use non-negative matrix factorization (NMF) [1–4]. It approximates the magnitude spectrogram of an observed music signal as the product of a basis matrix (spectral template vectors, each of which corresponds to a pitch) and an activation matrix (gain vectors, each of which is associated with a spectral template). NMF can be interpreted as statistical inference of a generative model that represents the process in which multiple pitches with time-invariant spectra are superimposed to generate an observed audio signal. There remain two major problems when we adopt it for multipitch estimation. First, the estimated activation matrix needs to be thresholded in post-processing to obtain a piano roll that indicates the existence of each pitch at each time unit (e.g., 16th-note length or time frame). An optimal threshold is different for each musical piece and is thus difficult to find. Second, relationships among two or more pitches are not considered in NMF, which may result in musically inappropriate estimations.

When humans manually transcribe music signals into musical scores, not only the audio reproducibility but also musical appropriateness of the scores is considered to avoid musically unnatural notes. Such musical appropriateness can be measured in accordance with a music theory (e.g., counterpoint theory and harmony theory). For instance, 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) and chords vary over time to form typical progressions.

Many studies have been conducted for estimating chords from musical scores [5–8]. If chord labels are given as clues for multipitch estimation, musically appropriate piano rolls is expected to be obtained. Typical chords and chord progressions, however, vary between music styles, e.g., the harmony theory of jazz music is different from that of classical music. It would thus be better to infer chords and their progressions adaptively for each musical piece. Since chords are

Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, Japan

Corresponding author:
Kazuyoshi Yoshii
Email: yoshii@kuis.kyoto-u.ac.jp

determined by note cooccurrences and *vice versa*, simultaneous estimation of chords and note cooccurrences is a chicken-and-egg problem. This indicates that it is appropriate to estimate chords and note cooccurrences in a unified framework.

In this paper, we propose a novel statistical method that discovers interdependent chords and pitches from music signals in an unsupervised manner (Fig. 1). We formulate a unified probabilistic generative model of a music spectrogram by integrating an *acoustic model* and a *language model* in the frame or tatum (16th-note-level beat) level, where the correct tatum times are assumed to be given in this paper. The acoustic model represents how the spectrogram is generated from a piano roll based on an extension of NMF with binary activations of pitches in the same way as [9]. The language model represents how the piano roll is generated from a chord sequence based on an autoregressive hidden Markov model (HMM) that considers the sequential dependencies of chords and pitches. In our previous study [10], we formulated only a frame-level unified model based on a standard HMM that considers only the sequential dependency of chords.

We then solve the inverse problem, i.e., given a music spectrogram, the whole model is inferred jointly. Since the acoustic and language models can be trained jointly in an unsupervised manner, the basis spectra of pitched instruments and typical note cooccurrences are learned directly from the observed music signal and all the latent variables (pitches and chords) are thus estimated jointly by using Gibbs sampling. Note that the language model can be trained in advance and the probabilities of typical note cooccurrences obtained from the training data are used as the parameters of the language model.

The major contribution of this study is to achieve grammar induction from music signals by integrating acoustic and language models. Both models are jointly learned in an unsupervised manner unlike a typical approach to automatic speech recognition (ASR). While ASR is based on a two-level hierarchy (word–spectrogram), our model has a three-level hierarchy (chord–pitch–spectrogram) by using an HMM instead of a Markov model (n-gram model) as a language model. We conducted comprehensive comparative experiments to evaluate the effectiveness of each component of the proposed unified model. Another important contribution is to release beat and chord annotations of the MAPS database [11] used for evaluation. Recently, ground-truth annotations of several musical elements (e.g., tempo, time signature, and key) for the MAP database were released by Ycart and Benetos [12]. Our annotations are complementary to their annotations.

The rest of the paper is organized as follows. Section II reviews related work on acoustic and language modeling. Section III explains the unified model based on acoustic and language models, and Section IV describes Bayesian inference of the model parameters and latent variables. Section V reports comparative evaluation using piano music data, and Section VI summarizes the paper.

II. RELATED WORK

This section reviews related work on acoustic modeling, language modeling, and integrated acoustic and language modeling for AMT.

A) Acoustic modeling

NMF and probabilistic latent component analysis (PLCA) are conventionally applied as the methods for spectrogram decomposition [1–4, 9, 13–18]. NMF approximates a non-negative matrix (a magnitude spectrogram) as the product of two non-negative matrices; bases (a set of spectral templates corresponding to different pitches or timbres) and activations (a set of gain vectors). Similarly, PLCA approximates a normalized spectrogram as a bivariate probability distribution and decomposes it into a series of spectral templates, pitches, instruments, and so on.

Smaragdis [19] proposed convolutive NMF that uses a time-frequency segment as a template. Virtanen *et al.* [3] reformulated bases as the product of sources corresponding to pitches and filters corresponding to timbres. This extension contributes to reducing the number of parameters and makes the estimation of bases more reliable, especially when different instruments play the same pitch. Vincent *et al.* [4] also extended NMF by forcing each basis to have harmonicity and spectral smoothness. Each pitch is represented as the sum of corresponding bases so that it adaptively fits the spectral envelope of a musical instrument in the observed music signal. O'Hanlon *et al.* [20] proposed group-sparse NMF that can represent the co-activity of bases. Using

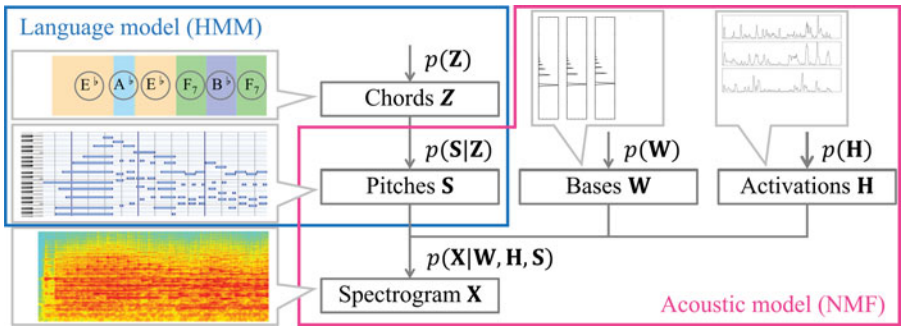

Fig. 1. A hierarchical generative model consisting of language and acoustic models that are linked through binary variables representing the existences of pitches.

group sparsity in addition to narrow bands proposed in [4], they let bases fit more adaptively to the observed signals. Cheng *et al.* [15] proposed an attack and decay model for piano transcription.

There have been some attempts to introduce prior knowledge into the NMF framework. Cemgil *et al.* [13] described Bayesian inference for NMF. Hoffman *et al.* [2] introduced a Bayesian non-parametric model called γ-process NMF to estimate an appropriate number of bases that are necessary to reconstruct the observation. Liang *et al.* [9] also proposed a Bayesian non-parametric extension called β-process NMF that multiplies a binary matrix (mask) to the activation matrix.

Deep learning techniques have recently been used for AMT. Nam *et al.* [21] used a deep belief network for learning latent representations of magnitude spectra and used support vector machines for judging the existence of each pitch. Boulanger-Lewandowski *et al.* [22] proposed a recurrent extension of the restricted Boltzmann machine and found that musically plausible transcriptions were obtained.

B) Language modeling

Some studies have attempted to computationally represent music theory. Hamanaka *et al.* [23] reformalized a systematized music theory called the generative theory of tonal music (GTTM) [24] and developed a method for estimating a tree that represents the structure of music called a time-span tree. Nakamura *et al.* [25] also reformalized the GTTM as a probabilistic context-free grammar. These methods enable automatic music parsing. Induction of harmony in an unsupervised manner has also been studied. Hu *et al.* [26] used latent Dirichlet allocation to determine the key of a musical piece from symbolic and audio music data based on the fact that the likelihood of the appearance of each note tends to be similar among musical pieces in the same key. This method enables the distribution of pitches in a certain key (key profile) to be obtained without using labeled training data.

Statistical methods of supervised chord recognition [5–8] are worth investigation for unsupervised music grammar induction. Rocher *et al.* [5] attempted chord recognition from symbolic music by constructing a directed graph of possible chords and then calculating the optimal path. Sheh *et al.* [6] used acoustic features called chroma vectors to estimate chords from music signals. They constructed an HMM whose latent variables are chord labels and whose observations are chroma vectors. Maruo *et al.* [7] proposed a method that uses NMF for extracting reliable chroma features. Since these methods require labeled training data, the concept of chords is required in advance.

C) Acoustic and language modeling

Multipitch estimation considering both acoustic features and music grammar has recently been studied. Raczyński *et al.* [27, 28] proposed a probabilistic pitch model based on a dynamic Bayesian network consisting of several submodels, each of which describes a different property of pitches. This model in combination with an NMF-based acoustic model performs better in multipitch estimation. Böck *et al.* [29] proposed a method for note onset transcription based on a recurrent neural network (RNN) with long short-term memory (LSTM) units that takes acoustic features as input and outputs a piano roll. Sigtia *et al.* [30] used an RNN as a language model. They integrated the RNN with a PLCA-based acoustic model so that the output of the RNN is treated as a prior for pitch activations. Holzapfel *et al.* [31] proposed a method that uses tatum information for multipitch estimation. Ycart *et al.* [32] used an LSTM network that takes a piano roll as an input and predict the next frame. The network is used for the post-processing of the piano roll estimated with the acoustic model proposed in [16]. In their study, tatum information was used to evaluate the note-prediction accuracy.

III. GENERATIVE MODELING

This section explains a generative model of a music spectrogram for estimating pitches and their typical cooccurrences (chords) from music signals. Our model consists of acoustic and language models 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 basis spectra and temporal activations of individual pitches. The language model represents the generative process of chord progressions and pitch locations from chords.

A) Problem specification

The goal of this study is to estimate a piano roll from a music signal played by pitched instruments. Let $\boldsymbol{X} = \{X_{ft}\}_{f,t=1}^{F,T}$ be the log-frequency magnitude spectrogram (e.g., constant-Q transform) of a music signal, where F is the number of frequency bins and T is the number of time frames. Let $\boldsymbol{S} = \{S_{kn}\}_{k,n=1}^{K,N}$ be a piano roll, where $S_{kn} \in \{0, 1\}$ indicates the existence of pitch k at the n-th time unit (tatum time or time frame) and K is the number of unique pitches. When we formulate a frame-level model without using tatum information, $T = N$ holds. When we formulate a tatum-level model. the tatum times are assumed to be given or estimated in advance. In addition, we aim to estimate a sequence of chords $\boldsymbol{Z} = \{z_n\}_{n=1}^{N}$ over N time units, where $z_n \in \{1, \ldots, I\}$ indicates a chord at the n-th time unit and I is the number of unique chords.

B) Acoustic modeling

We design a generative model of $\boldsymbol{X}$ inspired by a Bayesian extension of NMF with binary variables [9] (Fig. 2). The spectrogram $\boldsymbol{X} \in \mathbb{R}_+^{F\times T}$ is factorized into basis spectra $\boldsymbol{W} = \{\boldsymbol{W}^h \in \mathbb{R}_+^{K\times F}, \boldsymbol{W}^n \in \mathbb{R}_+^{1\times F}\}$ consisting of K harmonic spectra $\boldsymbol{W}^h$ and a noise spectrum $\boldsymbol{W}^n$, the corresponding temporal activations $\boldsymbol{H} = \{\boldsymbol{H}^h \in \mathbb{R}_+^{K\times T},$

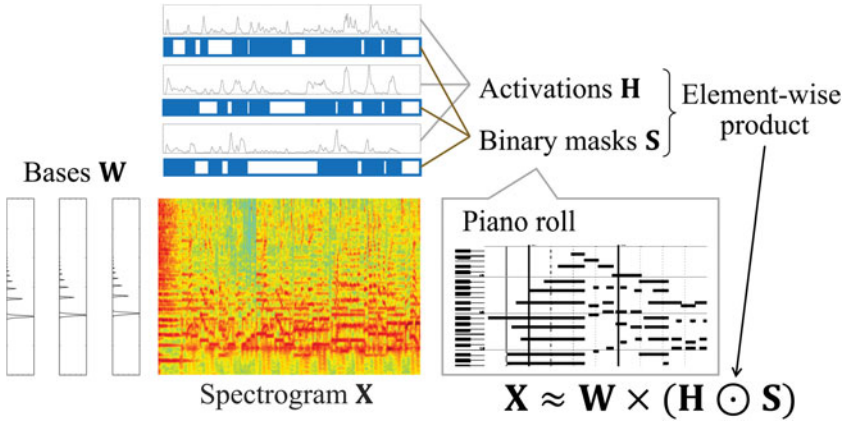

Fig. 2. An acoustic model based on a variant of NMF with binary variables indicating a piano roll.

$\boldsymbol{H}^n \in \mathbb{R}_+^{1\times T}\}$, and binary variables $\boldsymbol{S} \in \{0,1\}^{K\times N}$ as follows:

$$X_{ft}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{S} \sim \mathcal{P}\left(X_{ft}\middle|\sum_{k=1}^{K} W_{kf}^h H_{kt}^h S_{kn_t} + W_f^n H_t^n\right), \quad (1)$$

where $\mathcal{P}$ indicates a Poisson distribution, W_{kf}^h is the magnitude of harmonic basis k at frequency f, H_{kt}^h is its gain at frame t, and S_{kn_t} is a binary variable indicating whether basis k is activated at time n_t. Here, n_t is a time unit to which frame t belongs to ($n_t = t$ in a frame-level model). Similarly, W_f^n and H_t^n are defined for the noise component.

As proposed in [33], we assume that the harmonic structures of different pitches have shift-invariant relationships as follows:

$$W_{kf}^h = \overline{W}_{f_k}^h \ (1 \le k \le K), \quad (2)$$

where $\{\overline{W}_f^h\}_{f=1}^F$ is a template pattern shared by the K harmonic spectra, $f_k = f - (k-1)\Delta$ is a shifting interval, and Δ is the number of log-frequency bins corresponding to a semitone. If $f_k \le 0$, $W_{kf}^h = 0$. Although equation (2) is an excessively simplified model of real instrument sounds and the expressive capability of the acoustic model is limited, it contributes to automatically learning a harmonic template without explicitly imposing harmonic constraints (Fig. 3). We put a γ prior on $\overline{W}_f^h$ as follows:

$$\overline{W}_f^h \sim \mathcal{G}\left(a^h, b^h\right), \quad (3)$$

where a^h and b^h are shape and rate hyperparameters.

As proposed in [34], we put a γ chain prior on W_f^n to induce the spectral smoothness as follows:

$$\begin{cases} W_1^n \sim \mathcal{G}(\eta, \eta b^n/a^n), \\ G_{f-1}^n|W_{f-1}^n \sim \mathcal{G}\left(\eta, \eta W_{f-1}^n\right), \\ W_f^n|G_{f-1}^n \sim \mathcal{G}\left(\eta, \eta G_{f-1}^n\right), \end{cases} \quad (4)$$

where η is a hyperparameter adjusting the degree of smoothness and G_f^n is an auxiliary variable forcing W_{f-1}^n to be positively correlated with W_f^n. Since $\mathbb{E}_{prior}[G_{f-1}^n] = 1/W_{f-1}^n$ and $\mathbb{E}_{prior}[W_f^n] = 1/G_{f-1}^n$, we can roughly say $\mathbb{E}_{prior}[W_f^n] \approx \mathbb{E}_{prior}[W_{f-1}^n]$ (Fig. 3).

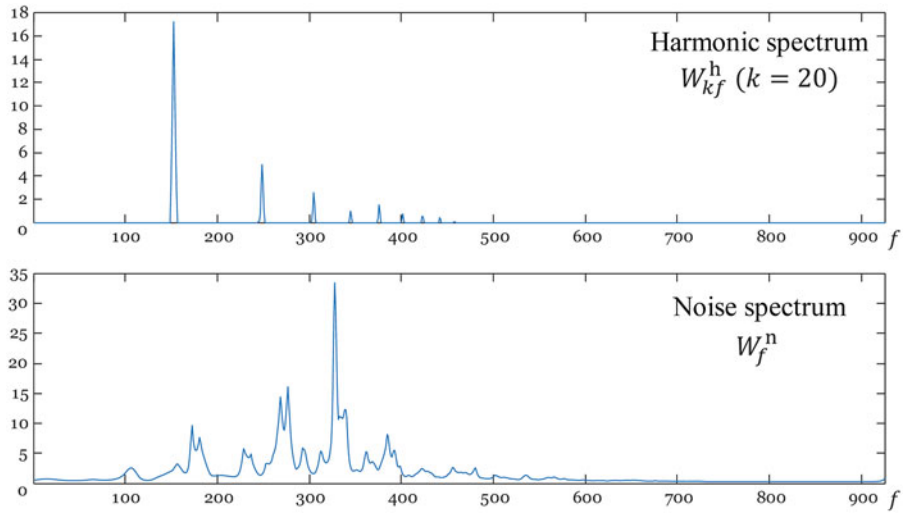

Fig. 3. Harmonic and noise spectra learned from data.

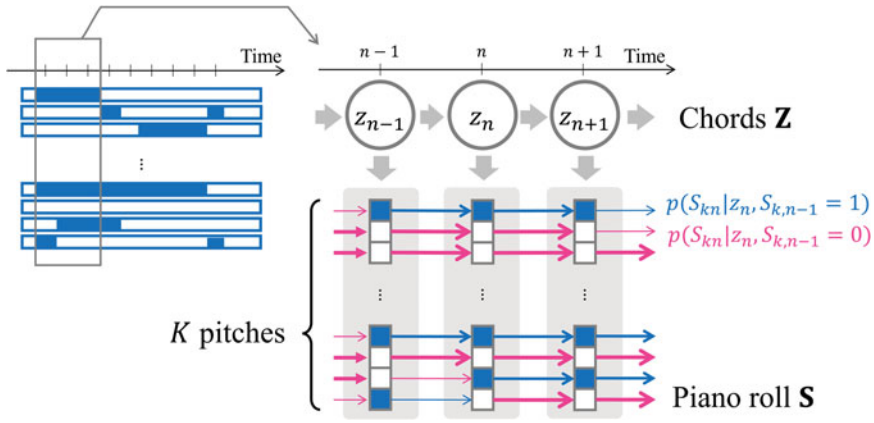

Fig. 4. A language model based on an autoregressive HMM that emits sequentially dependent binary variables.

We put γ priors on the activations $\boldsymbol{H}$ as follows:

$$H_{kt}^h \sim \mathcal{G}\left(c^h, d^h\right), \tag{5}$$

$$H_t^n \sim \mathcal{G}\left(c^n, d^n\right), \tag{6}$$

where c^h, c^n, d^h, and d^n are hyperparameters.

C) Language modeling

We propose an HMM that has latent variables (chords) $\boldsymbol{Z}$ and emits binary variables (pitches) $\boldsymbol{S}$, which cannot be observed in reality, as follows (Fig. 4):

$$z_1|\boldsymbol{\phi} \sim Categorical(\boldsymbol{\phi}), \tag{7}$$

$$z_n|z_{n-1}, \boldsymbol{\psi} \sim Categorical(\boldsymbol{\psi}_{z_{n-1}}), \tag{8}$$

$$S_{kn}|z_n, \boldsymbol{\pi} \sim Bernoulli(\pi_{z_n,k}), \tag{9}$$

where $\boldsymbol{\phi} \in \mathbb{R}_+^I$ is a set of initial probabilities, $\boldsymbol{\psi}_i \in \mathbb{R}_+^I$ is a set of transition probabilities from chord i, and π_{ik} indicates the emission probability of pitch k from chord i. In this paper, we focus on only the emission probabilities of the 12 pitch classes (C, C♯, ..., B, $m = 0, \ldots, 11$), which are copied to all octaves covering the K pitches. Let $\overline{\pi}_{jm}$ be the emission probability of pitch class m from chord type j (major or minor, $j = 0, 1$). The emission probabilities from chords of the same type are assumed to have circular-shifting relationships as follows:

$$\pi_{ik} = \overline{\pi}_{type(i),mod(class(k)-root(i),12)}, \tag{10}$$

where $type(i) \in \{0, 1\}$ and $root(i) \in \{0, \ldots, 11\}$ are the type and root note of chord i, respectively, and $class(k) \in \{0, \ldots, 11\}$ is the pitch class of pitch k. We put conjugate priors on those parameters as follows:

$$\boldsymbol{\phi} \sim Dir(\boldsymbol{u}), \tag{11}$$

$$\boldsymbol{\psi}_i \sim Dir(\boldsymbol{v}_i), \tag{12}$$

$$\overline{\pi}_{jm} \sim \beta(e, f), \tag{13}$$

where $\boldsymbol{u} \in \mathbb{R}_+^I$, $\boldsymbol{v}_i \in \mathbb{R}_+^I$, e, and f are hyperparameters.

This HMM can be extended in an autoregressive manner by incorporating the sequential dependency (smoothness) of binary variables of each pitch. More specifically, equation (9) can be extended as follows [35]:

$$S_{kn}|z_n, S_{k,n-1}, \boldsymbol{\pi} \sim Bernoulli\left(\pi_{z_n,k}^{(S_{k,n-1})}\right), \tag{14}$$

where $\pi_{z_n,k}^{(S_{k,n-1})}$ indicates the emission probability of pitch k from chord z_n at time unit n when the same pitch k is activated ($S_{k,n-1} = 1$) or not activated ($S_{k,n-1} = 0$) at the previous time unit $n - 1$. Instead of equation (13), we consider two types of emission probabilities as follows:

$$\overline{\pi}_{jm}^{(0)} \sim \beta(e^{(0)}, f^{(0)}), \tag{15}$$

$$\overline{\pi}_{jm}^{(1)} \sim \beta(e^{(1)}, f^{(1)}), \tag{16}$$

where $e^{(0)}$, $f^{(0)}$, $e^{(1)}$, and $f^{(1)}$ are hyperparameters. The circular-shifting relationships between $\pi_{ik}^{(0)}$ and $\overline{\pi}_{jm}^{(0)}$ and that between $\pi_{ik}^{(1)}$ and $\overline{\pi}_{jm}^{(1)}$ are defined in the same way as equation (10). The self-transitions (i.e., $S_{k,n-1} = S_{k,n} = 0$ and $S_{k,n-1} = S_{k,n} = 1$) are more likely to occur when $e^{(0)} \gg f^{(0)}$ and $e^{(1)} \ll f^{(1)}$. This contributes to reducing spurious musical notes that tend to have very short durations. We evaluated the standard HMM given by equations (9) and (13) and the autoregressive HMM given by equations (14)–(16) in Section V.

IV. POSTERIOR INFERENCE

This section describes AMT for the observed data $\boldsymbol{X}$. We explain Bayesian inference of the proposed model and then describe how to put emphasis on the language model. In addition, we describe how to pre-train the language model.

A) Bayesian inference

Given the observation $\boldsymbol{X}$, we aim to calculate the posterior distribution $p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{S}, \boldsymbol{\Theta}|\boldsymbol{X})$, where $\boldsymbol{\Theta} = \{\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\pi}\}$. We use Gibbs sampling to approximate the analytically intractable posterior distribution (Algorithm 1). The piano

Algorithm 1 Posterior inference

Require: Hyperparameters of acoustic model:
  Set γ priors on basis spectra $\boldsymbol{W}$
    by specifying a^h, b^h (equation (3)), a^n, b^n, η
    (equation (4))
  Set γ priors on temporal activations $\boldsymbol{H}$
    by specifying c^h, d^h (equation (5)), c^n, d^n (equation (6))
Require: Hyperparameters of language model:
  Set Dirichlet prior on initial probabilities $\boldsymbol{\phi}$
    by specifying $\boldsymbol{u}$ (equation (11))
  Set Dirichlet priors on transition probabilities $\boldsymbol{\psi}$
    by specifying $\boldsymbol{v}_i$ (equation (12))
  Set β priors on emission probabilities $\boldsymbol{\pi}$
    by specifying e, f (equation (13))
    or $e^{(0)}$, $f^{(0)}$, $e^{(1)}$, $f^{(1)}$ (equations (15) and (16))
  procedure GibbsSampling
    Initialize piano roll $\boldsymbol{S}$
    Initialize NMF parameters $\boldsymbol{W}$ and $\boldsymbol{H}$
    Initialize chord sequence $\boldsymbol{Z}$
    Initialize HMM parameters $\boldsymbol{\Theta} = \{\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\pi}\}$
    loop
      Update $\boldsymbol{S}$ (equation (17))
      Update $\boldsymbol{W}$ and $\boldsymbol{H}$ (equations (20), (23), and (28))
      Update $\boldsymbol{Z}$ (equation (31))
      Update $\boldsymbol{\Theta}$ (equations (37) and (45))
    end loop
  end procedure

roll $\boldsymbol{S}$ is estimated by using the current acoustic and language models, which are then updated independently by using the current $\boldsymbol{S}$. These steps are iterated until approximate convergence. Finally, a sequence of chords $\boldsymbol{Z}$ is estimated using the Viterbi algorithm and then $\boldsymbol{S}$ is determined using the maximum-likelihood parameters of the unified model.

1) Updating piano roll

The piano roll $\boldsymbol{S}$ is sampled in an element-wise manner from a conditional posterior distribution (Bernoulli distribution) obtained by integrating the acoustic model with the language model as follows:

$$\begin{aligned} &p(S_{kn}|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{S}_{\neg kn}, \boldsymbol{\Theta}) \\ &\quad \propto p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S})\, p(S_{kn}|z_n, \boldsymbol{S}_{\neg kn}, \boldsymbol{\pi}), \end{aligned} \tag{17}$$

where a notation $\boldsymbol{\Omega}_{\neg i}$ indicates a set of all elements of $\boldsymbol{\Omega}$ except for the i-th element, the first term (likelihood, acoustic model) is given by equation (1), and the second term (prior, language model) is given by equation (14) as follows:

$$\begin{aligned} &p(S_{kn}|z_n, \boldsymbol{S}_{\neg kn}, \boldsymbol{\pi}) \\ &\quad \propto \left(\pi_{z_n,k}^{(S_{k,n-1})}\right)^{S_{kn}} \left(1 - \pi_{z_n,k}^{(S_{k,n-1})}\right)^{1-S_{kn}} \\ &\qquad \times \left(\pi_{z_{n+1},k}^{(S_{k,n})}\right)^{S_{k,n+1}} \left(1 - \pi_{z_{n+1},k}^{(S_{k,n})}\right)^{1-S_{k,n+1}}. \end{aligned} \tag{18}$$

When the sequential dependency of $\boldsymbol{S}$ is not considered, equation (18) is simplified as follows:

$$p(S_{kn}|z_n, \boldsymbol{S}_{\neg kn}, \boldsymbol{\pi}) \propto \left(\pi_{z_n,k}\right)^{S_{kn}} \left(1 - \pi_{z_n,k}\right)^{1-S_{kn}}. \tag{19}$$

2) Updating acoustic model

The parameters $\boldsymbol{W}$ and $\boldsymbol{H}$ of the acoustic model are sampled using Gibbs sampling in the same way as [9]. Note that $\boldsymbol{W}^h$ and $\boldsymbol{H}$ have γ priors and $\boldsymbol{W}^n$ have γ chain priors. Because of the conjugacy, we can easily calculate the γ posterior of each variable conditioned on the other variables and the binary variables.

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

$$p(\overline{\boldsymbol{W}}^h|\boldsymbol{X}, \boldsymbol{W}^n, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{S}) \propto p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S})\, p(\overline{\boldsymbol{W}}^h), \tag{20}$$

where the first term (likelihood) is given by equation (1) and the second term (prior) is given by equation (3). More specifically, we obtain

$$\begin{aligned} \overline{W}_f^h \sim \mathcal{G}\Bigg(&\sum_{k=1}^{K}\sum_{t=1}^{T} X_{\tilde{f}_k t}\lambda_{\tilde{f}_k tk}^h + a^h, \\ &\sum_{k=1}^{K}\sum_{t=1}^{T} H_{kt}^h S_{kn_t} + b^h \Bigg), \end{aligned} \tag{21}$$

where $\tilde{f}_k = f + (k-1)\Delta$ (if $\tilde{f}_k > F$, $X_{\tilde{f}_k t} = 0$) and λ_{ftk}^h is an auxiliary variable obtained by using the previous samples of $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{S}$ as follows:

$$\lambda_{ftk}^h = \frac{W_{kf}^h H_{kt}^h S_{kn_t}}{\sum_{k'} W_{k'f}^h H_{k't}^h S_{k'n_t} + W_f^n H_t^n}. \tag{22}$$

Since $\boldsymbol{W}^n$ and $\boldsymbol{G}^n$ are interdependent in equation (4), $\boldsymbol{W}^n$ and $\boldsymbol{G}^n$ are sampled alternately as follows:

$$\begin{aligned} &p(\boldsymbol{W}^n|\boldsymbol{X}, \boldsymbol{W}^h, \boldsymbol{G}^n, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{S}) \\ &\quad \propto p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S})\, p(\boldsymbol{W}^n, \boldsymbol{G}^n), \end{aligned} \tag{23}$$

$$p(\boldsymbol{G}^n|\boldsymbol{X}, \boldsymbol{W}^h, \boldsymbol{W}^n, \boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{S}) \propto p(\boldsymbol{W}^n, \boldsymbol{G}^n), \tag{24}$$

where the first and second terms of equation (23) are given by equations (1) and (4), respectively. More specifically, we obtain

$$\begin{aligned} W_f^n \sim \mathcal{G}\Bigg(&\sum_{t=1}^{T} X_{ft}\lambda_{ft}^n + \eta, \\ &\sum_{t=1}^{T} H_t^n + \eta\left(G_{f+1}^n + G_f^n\right) \Bigg), \end{aligned} \tag{25}$$

$$G_f^n \sim \mathcal{G}\left(\eta, \eta\left(W_f^n + W_{f-1}^n\right)\right), \tag{26}$$

where λ_{ft}^n is an auxiliary variable given by

$$\lambda_{ft}^n = \frac{W_f^n H_t^n}{\sum_{k'} W_{k'f}^h H_{k't}^h S_{k'n_t} + W_f^n H_t^n}. \tag{27}$$

The conditional posterior distribution of $\boldsymbol{H}$ is given by

$$p(\boldsymbol{H}|\boldsymbol{X},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{S},\boldsymbol{\Theta}) \propto p(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{S})p(\boldsymbol{H}), \tag{28}$$

where the first term (likelihood) is given by equation (1) and the second term (prior) is given by equations (3) and (6). More specifically, we obtain

$$H^h_{kt} \sim \mathcal{G}\left(\sum_{f=1}^{F} X_{ft}\lambda^h_{ftk} + c^h,\ S_{kn_t}\sum_{f=1}^{F} W^h_{kf} + d^h\right), \tag{29}$$

$$H^n_t \sim \mathcal{G}\left(\sum_{f=1}^{F} X_{ft}\lambda^n_{ft} + c^n,\ \sum_{f=1}^{F} W^n_f + d^n\right). \tag{30}$$

3) Updating chord sequence

The latent variables $\boldsymbol{Z}$ can be updated efficiently by using a forward filtering-backward sampling algorithm, which is a stochastic version of the forward-backward algorithm (Baum–Welch algorithm). The conditional posterior distribution of $\boldsymbol{Z}$ is given by

$$p(\boldsymbol{Z}|\boldsymbol{S},\boldsymbol{\Theta}) \propto p(\boldsymbol{S}|\boldsymbol{Z},\boldsymbol{\pi})p(\boldsymbol{Z}|\boldsymbol{\phi},\boldsymbol{\psi}), \tag{31}$$

where the first term is given by equation (9) and the second term is given by equations (7) and (8). Let a Matlab-like notation $\boldsymbol{s}_{1:n}$ denote $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n\}$, where $\boldsymbol{s}_n = [S_{n1},\ldots,S_{nK}]^T$. Henceforth, we often omit the dependency on $\boldsymbol{\Theta}$ for brevity ($\boldsymbol{\Theta}$ is assumed to be given for estimating $\boldsymbol{Z}$). As in a standard HMM, a forward message $\alpha(z_n) = p(\boldsymbol{s}_{1:n}, z_n)$ can be calculated recursively as follows:

$$\alpha(z_1) = p(z_1)p(\boldsymbol{s}_1|z_1), \tag{32}$$

$$\alpha(z_n) = p(\boldsymbol{s}_n|z_n,\boldsymbol{s}_{n-1})\sum_{z_{n-1}} p(z_n|z_{n-1})\alpha(z_{n-1}), \tag{33}$$

where $p(z_1) = \phi_{z_1}$, $p(z_n|z_{n-1}) = \psi_{z_{n-1},z_n}$, and $p(\boldsymbol{s}_1|z_1)$ and $p(\boldsymbol{s}_n|z_n,\boldsymbol{s}_{n-1})$ are given by equation (14) or equation (9).

After calculating the forward messages, we perform the backward sampling as follows:

$$p(\boldsymbol{Z}|\boldsymbol{S}) = p(z_N|\boldsymbol{S})\prod_{n=1}^{N-1} p(z_n|\boldsymbol{S},z_{n+1:N}), \tag{34}$$

More specifically, the last latent variable z_N is sampled as follows:

$$z_N \sim p(z_N|\boldsymbol{S}) \propto \alpha(z_N). \tag{35}$$

The other latent variables $z_{1:N-1}$ are then sampled recursively in the reverse order as follows:

$$z_n \sim p(z_n|\boldsymbol{S},z_{n+1:N}) \propto p(z_{n+1}|z_n)\alpha(z_n). \tag{36}$$

4) Updating language model

Using the Bayes' rule, the posterior distribution of the emission probabilities $\overline{\boldsymbol{\pi}}$ is given by

$$p(\overline{\boldsymbol{\pi}}|\boldsymbol{S},\boldsymbol{Z}) \propto p(\boldsymbol{S}|\boldsymbol{Z},\overline{\boldsymbol{\pi}})p(\overline{\boldsymbol{\pi}}), \tag{37}$$

where the first term (likelihood) is given by equation (9) and the second term (prior) is give by equation (15) and equation (16). More specifically, we obtain

$$\overline{\pi}^{(0)}_{jm} \sim \beta\left(e^{(0)} + r^{(01)}_{jm}, f^{(0)} + r^{(00)}_{jm}\right), \tag{38}$$

$$\overline{\pi}^{(1)}_{jm} \sim \beta\left(e^{(1)} + r^{(11)}_{jm}, f^{(1)} + r^{(10)}_{jm}\right), \tag{39}$$

where $r^{(00)}_{jm}$, $r^{(01)}_{jm}$, $r^{(10)}_{jm}$, and $r^{(11)}_{jm}$ are count data given by

$$r^{(00)}_{jm} = \sum_{n\in A_j}\sum_{k\in B_{nm}} (1-S_{k,n-1})(1-S_{kn}), \tag{40}$$

$$r^{(01)}_{jm} = \sum_{n\in A_j}\sum_{k\in B_{nm}} (1-S_{k,n-1})S_{kn}, \tag{41}$$

$$r^{(10)}_{jm} = \sum_{n\in A_j}\sum_{k\in B_{nm}} S_{k,n-1}(1-S_{kn}), \tag{42}$$

$$r^{(11)}_{jm} = \sum_{n\in A_j}\sum_{k\in B_{nm}} S_{k,n-1}S_{kn}, \tag{43}$$

where A_j and B_{nm} are subsets of indices given by $A_j = \{n|type(z_n) = j\}$ and $B_{nm} = \{k|mod(class(k) - root(z_n), 12) = m\}$. When the sequential dependency of $\boldsymbol{S}$ is not considered, we obtain

$$\overline{\pi}_{jm} \sim \beta\left(e + r^{(00)}_{jm} + r^{(10)}_{jm}, f + r^{(01)}_{jm} + r^{(11)}_{jm}\right). \tag{44}$$

The posterior distributions of the initial probabilities $\boldsymbol{\phi}$ and the transition probabilities $\boldsymbol{\psi}$ are given by

$$p(\boldsymbol{\phi},\boldsymbol{\psi}|\boldsymbol{S},\boldsymbol{Z}) \propto p(\boldsymbol{Z}|\boldsymbol{\phi},\boldsymbol{\psi})p(\boldsymbol{\phi})p(\boldsymbol{\psi}), \tag{45}$$

where the first term is given by equations (7) and (8), the second term is given by equation (11), and the third term is given by equation (12). More specifically, we obtain

$$\boldsymbol{\phi} \sim Dir\left(\mathbf{1}_I + \boldsymbol{e}_{z_1}\right), \tag{46}$$

$$\boldsymbol{\psi}_i \sim Dir\left(\boldsymbol{v}_i + \boldsymbol{u}_i\right), \tag{47}$$

where $\boldsymbol{e}_i$ is the unit vector whose i-th element is 1 and $\boldsymbol{u}_i$ is the I-dimensional vector whose j-th element indicates the number of transitions from state i to state j.

B) Weighted integration

In naive integration of the language model and acoustic models, the language model does not effectively affect the posterior distribution of piano roll $\boldsymbol{S}$, i.e., musically inappropriate allocation of musical notes is not given a large penalty. To balance the impact of the language model with

that of the acoustic model, we introduce a weighting factor α as in ASR, i.e., equation (18) is replaced with

$$p(S_{kn}|z_n, \boldsymbol{S}_{\neg kn}, \boldsymbol{\pi}) \propto \left(\left(\pi_{z_n,k}^{(S_{k,n-1})} \right)^{S_{kn}} \left(1 - \pi_{z_n,k}^{(S_{k,n-1})} \right)^{1-S_{kn}} \times \left(\pi_{z_{n+1},k}^{(S_{k,n})} \right)^{S_{k,n+1}} \left(1 - \pi_{z_{n+1},k}^{(S_{k,n})} \right)^{1-S_{k,n+1}} \right)^{\alpha D_n}, \quad (48)$$

where D_n indicates the number of time frames in time unit n. When the sequential dependency of $\boldsymbol{S}$ is not considered, equation (19) is replaced with

$$p(S_{kn}|z_n, \boldsymbol{S}_{\neg kn}, \boldsymbol{\pi}) \propto \left(\left(\pi_{z_n,k} \right)^{S_{kn}} \left(1 - \pi_{z_n,k} \right)^{1-S_{kn}} \right)^{\alpha D_n}. \quad (49)$$

We empirically investigated the effect of these modifications (see Section V).

C) Prior training

The language model can be trained in advance from existing piano rolls (musical scores) even if no chord annotations are available. Here we assume that there a single piano roll $\hat{\boldsymbol{S}}$ is given as training data for simplicity because it is straightforward to deal with multiple piano rolls. The underlying chords $\hat{\boldsymbol{Z}}$ and the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\pi}\}$ can be estimated from $\hat{\boldsymbol{S}}$ instead of $\boldsymbol{S}$ as in Section IV-A-3. After using the Gibbs sampling, we determine $\hat{\boldsymbol{Z}}$ by using the Viterbi algorithm and calculate the posterior distributions of $\overline{\boldsymbol{\pi}}$ based on the estimate of $\hat{\boldsymbol{Z}}$ according to equations (38) and (39), which is then used as a prior distribution of $\overline{\boldsymbol{\pi}}$ instead of equations (15) and (16). This is a strong advantage of Bayesian formulation. Since the chord transitions are different for each musical piece, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are trained in an unsupervised manner.

V. EVALUATION

We evaluated the performance of the proposed method for AMT. First, we conducted a preliminary experiment to confirm that the language model can learn chord progressions and typical cooccurrences of pitches from piano rolls in an unsupervised manner. Next, we evaluated the performances of multipitch estimation obtained by using different language models and those obtained by the pre-trained and unsupervised models. Finally, we compared the proposed model with several unsupervised models.

A) Experimental conditions

We used 30 classical piano pieces labeled as 'ENSTDkCl' selected from the MAPS database [11]. An audio signal of 30 sec was extracted from the beginning of each piece. The magnitude spectrogram of size $F = 926$ and $T = 3000$ was obtained using variable-Q transform [36], where the number of frequency bins in one octave was set to 96. A harmonic and percussive source separation method [37] was used for suppressing non-harmonic components. All hyperparameters were determined empirically for maximizing the performance as described below.

We considered $K = 84$ unique pitches (MIDI note numbers 21–104) and $I = 24$ unique chords. We manually made tatum and chord annotations on 16th-note-level grids.[1] Since noise components were assumed to be smaller than harmonic components, the hyperparameters of $\boldsymbol{W}^h$ and $\boldsymbol{W}^n$ were set as $a^h = b^h = 1$, $a^n = 2$, and $b^n = 4$ such that $\mathbb{E}[\overline{W}_f^h] = 1$ and $\mathbb{E}[W_f^n] = 0.5$. The hyperparameters of $\boldsymbol{H}^h$ were set as $c^h = 10$ and $d^h = 10$ to favor non-zero gains. This was effective to avoid $S_{kn_t} = 1$ when H_{kt} takes almost zero. The hyperparameters of $\boldsymbol{H}^n$, on the other hand, were set as $c^n = 5$ and $d^n = 5$ to allow H_t^n to take almost zero. The hyperparameters of $\boldsymbol{\pi}$ were set as $e = e^{(1)} = e^{(0)} = 10^{-9}$, $f = f^{(1)} = f^{(0)} = 1$ to make a binary matrix $\boldsymbol{S}$ sparse. The hyperparameter η was empirically determined as $\eta = 30000$.

We tested two kinds of time resolutions for the language model, i.e., a *frame-level model* with a time resolution of 10 ms and a *tatum-level model* defined on a 16th-note-level grid. The hyperparameter of the initial probabilities $\boldsymbol{\phi}$ was set as $\boldsymbol{u} = [1, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^I$. In the frame-level model, the self-transition probability of each chord i was set as $\psi_{ii} = 0.99$ to favor temporal continuity and the transition probabilities from chord i to the other 23 chords were assumed to follow a Dirichlet distribution, $100\boldsymbol{\psi}_{i,\neg i} \sim Dir(\mathbf{1}_{I-1})$. In the tatum-level model, the hyperparameter of the transition probabilities $\boldsymbol{\psi}$ was set as $\boldsymbol{v}_i = [0.2, \ldots, 1, \ldots, 0.2]^{\mathrm{T}} \in \mathbb{R}^I$, where only the i-th dimension takes 1. The weighting factor of the language model, which has a strong impact on the performance, was empirically set to $\alpha = 12.5$ unless otherwise noted.

B) Chord estimation for piano rolls

We investigated whether chord progressions and typical cooccurrences of pitches (chords) can be learned from a piano roll obtained by concatenating the ground-truth piano rolls of the 30 pieces. The size of the matrix used as an input was thus $84 \times \sum_{i=1}^{30} N_i$, where N_i indicates the number of time frames or tatum times in the beginning 30 s of the i-th musical piece. We measured the performance of chord estimation as the ratio of the number of correctly estimated time frames or tatum times to the total number of those with major, minor, dominant 7th, and minor 7th chords. Dominant 7th was treated as a major chord, minor 7th as a minor chord, and the other chords were ignored. Since chords were estimated in an unsupervised manner, the estimated states were associated with chord labels to maximize the performance while conserving circular-shifting relationships.

[1] The beat and chord annotations used for evaluation are available on http://sap.ist.i.kyoto-u.ac.jp/members/yoshii/annotations/MAPS_beats.zip, http://sap.ist.i.kyoto-u.ac.jp/members/yoshii/annotations/MAPS_chords.zip

Table 1. Accuracy of unsupervised chord estimation.

Pitch emission	Frame-level model	Tatum-level model
Independent	58.9%	66.5%
Markov	42.7%	50.3%

As shown in Table 1, the accuracy of unsupervised chord estimation was around 60% and the tatum-level model outperformed the frame-level model. As shown in Fig. 5, chord structures (emission probabilities $\overline{\boldsymbol{\pi}}_0$ and $\overline{\boldsymbol{\pi}}_1$) corresponding to major and minor chords were learned when all the elements of $\boldsymbol{S}$ were assumed to be independent. When the sequential dependency of $\boldsymbol{S}$ was considered, the emission probabilities $\overline{\boldsymbol{\pi}}_0^{(0)}$, $\overline{\boldsymbol{\pi}}_1^{(0)}$, $\overline{\boldsymbol{\pi}}_0^{(1)}$, and $\overline{\boldsymbol{\pi}}_1^{(1)}$ were strongly affected by the previous binary variables, as shown in Fig. 6. Interestingly, when the previous binary variables were 0, typical pitch structures corresponding to a major chord and the diatonic scale were learned. This implies that a musical scale is more focused on than a chord when a new sound occurs. When the previous binary variable of a pitch was 1, the model prefers to continuously activate the pitch regardless of its pitch class because musical sounds usually continue for several time units.

C) Multipitch estimation for music signals

We evaluated the performance of multipitch estimation in the frame level in terms of the recall rate, precision rate, and F-measure defined as

$$\mathcal{R} = \frac{\sum_{t=1}^{T} c_t}{\sum_{t=1}^{T} r_t}, \quad \mathcal{P} = \frac{\sum_{t=1}^{T} c_t}{\sum_{t=1}^{T} e_t}, \quad \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}}, \tag{50}$$

where r_t, e_t, and c_t indicate the numbers of ground-truth, estimated, and correct pitches at time frame t, respectively. The tatum-level measures are defined similarly.

In addition, we measured the note-onset F-measure $\mathcal{F}_{on}$ [38] defined as follows:

$$\mathcal{R}_{on} = \frac{N_{det}}{N_{gt}}, \ \mathcal{P}_{on} = \frac{N_{cor}}{N_{est}}, \mathcal{F}_{on} = \frac{2\mathcal{R}_{on}\mathcal{P}_{on}}{\mathcal{R}_{on}+\mathcal{P}_{on}}, \tag{51}$$

where N_{det} is the number of musical notes that were included both in the ground-truth data and in output of the model, N_{gt} is the number of musical notes in the ground-truth data, N_{cor} is the number of musical notes regarded as correct in the estimated notes, and N_{est} is the number of musical notes in the output. For the frame-level model, an estimated note was regarded as correct if its pitch matched a ground-truth pitch and its onset was within 50 ms of

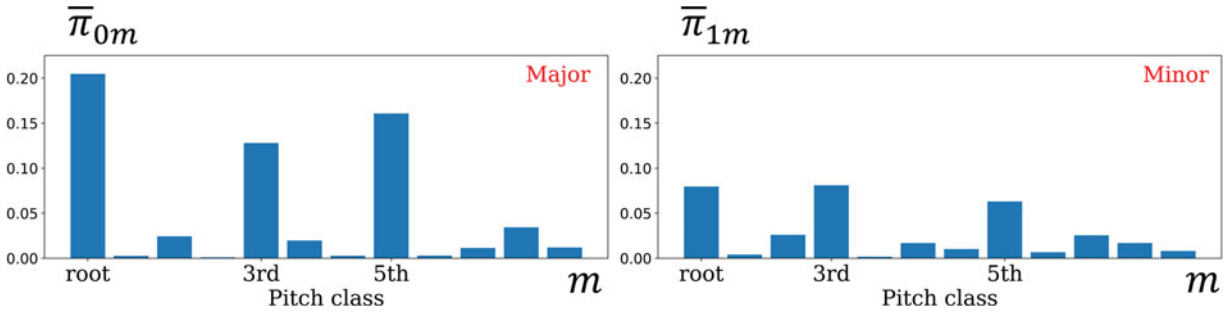

Fig. 5. The emission probabilities $\overline{\pi}$ obtained by the tatum-level model assuming the independence of $\boldsymbol{S}$.

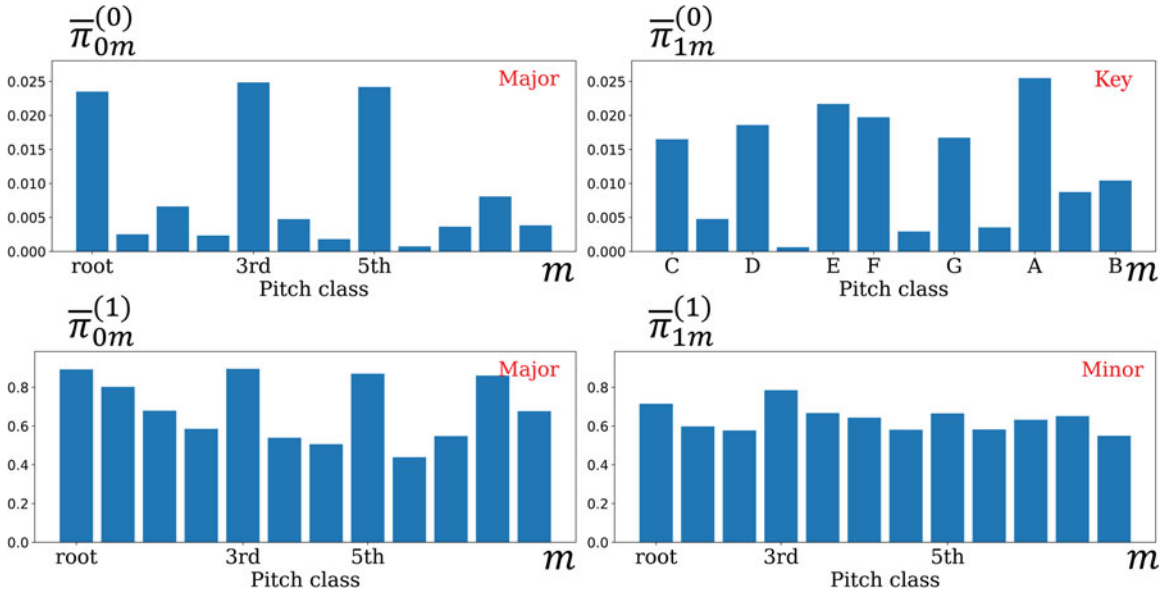

Fig. 6. The emission probabilities $\overline{\pi}$ obtained by the tatum-level model assuming the sequential dependency of $\boldsymbol{S}$.

the ground-truth onset. A musical note in the ground-truth data was regarded as detected if its pitch matched an estimated pitch and its onset was within 50 ms of an estimated note onset. For the tatum-level model, an estimated note was regarded as correct only if both its pitch and onset matched the ground-truth ones. A musical note in the ground-truth data was regarded as detected if both its pitch and its onset matched the estimated ones.

1) Evaluation of language modeling

We evaluated the effectiveness of each component of the language model by testing different priors on piano roll $\boldsymbol{S}$. More specifically, we compared the performances of the following five conditions:

(a) Uniform model: The emission probabilities $\boldsymbol{\pi}$ were fixed to 0.0625.
(b) Sparse model [9]: The emission probabilities $\boldsymbol{\pi}$ were assumed to be independent and were given sparse prior distributions $\pi_{ik} \sim \beta(10^{-9}, 1)$.
(c) Key-aware model: The emission probabilities $\boldsymbol{\pi}$ were estimated by fixing the latent variables $\boldsymbol{Z}$ to the same value, i.e., $z_1 = \cdots = z_N = 1$. In this case, the emission probabilities $\overline{\boldsymbol{\pi}}_1$ of the 12 pitch classes were expected to indicate the key profile of a target piece.
(d) Chord-aware model (HMM): Both $\boldsymbol{Z}$ and $\boldsymbol{\pi}$ were estimated by equations (36) and (44), respectively, without considering the sequential dependency of each pitch.
(e) Chord-aware Markov model (autoregressive HMM): Both $\boldsymbol{Z}$ and $\boldsymbol{\pi}$ were estimated by equation (36) and equations (38) and (39), respectively, based on the sequential dependency of each pitch.

We further examined the impact of the weighting factor α by testing $\alpha = 1$ and $\alpha = 12.5$ under the conditions (b) and (d). The performances were measured in the frame or tatum level. To evaluate the frame-level model in the tatum level, the existence of each pitch in each tatum interval was determined by taking the majority of binary variables in the interval. It was straightforward to evaluate the tatum-level model in the frame level.

As shown in Tables 2 and 3, the chord-aware model that does not consider sequential dependency of pitches performed best (69.8% without tatum information and 71.3% with tatum information). Note that the tatum information was not used in the frame-level evaluation of the frame-level model while it was used under the other conditions for estimation and/or evaluation. The F-measure obtained by the frame-level model was improved for 28 out of the 30 pieces (58.8% $\rightarrow$ 69.8%) by jointly estimating chords and pitches, even when the language and acoustic models were equally considered (56.8 $\rightarrow$ 57.6%). If the weighting factor α was increased from $\alpha = 1$ to $\alpha = 12.5$, the F-measure was significantly improved from 57.6 to 69.8%, even when the only key profile was learned (69.3%). This indicates the effectiveness of the language model weighting as discussed in Section IV-B. Introducing chord transitions further improved the F-measure from 69.3 to 69.8%.

Examples of estimated piano rolls are shown in Fig. 7. The F-measure obtained by the sparse model (Fig. 7(a)) was improved by 1.7 pts by estimating chords (Fig. 7(b)), and was further improved by 3.1 pts by emphasizing the language model (Fig. 7(c)). The sequential-dependency modeling of a piano roll, however, degraded the performance (Fig. 7(d)). As shown in Fig. 9, the emission probabilities estimated by the chord-aware Markov model indicate that the language model tends to focus on temporal continuity of pitches instead of learning typical note cooccurrences as chords. To solve this problem, the language model should be improved

Table 2. Experimental results of multipitch analysis based on the frame-level model for 30 piano pieces labeled as `ENSTDkCl`.

Language model (prior distribution on $\boldsymbol{S}$)	Frame-level evaluation				Tatum-level evaluation			
	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{F}_{on}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{F}_{on}$
Uniform model	88.7	38.8	53.0	9.7	89.3	40.8	54.9	30.8
Sparse model ($\alpha = 12.5$)	77.5	49.6	58.8	37.0	78.4	50.6	59.8	55.2
Key-aware model ($\alpha = 12.5$)	70.4	71.3	69.3	**50.6**	71.6	73.1	70.8	**64.8**
Chord-aware model ($\alpha = 12.5$)	73.7	69.0	**69.8**	49.4	75.1	70.6	**71.3**	64.1
Chord-aware Markov model ($\alpha = 12.5$)	87.9	43.8	57.3	22.0	88.5	44.8	58.2	37.0
Sparse model ($\alpha = 1$)	87.6	43.2	56.8	13.6	88.3	45.0	58.5	35.8
Chord-aware model ($\alpha = 1$)	87.5	44.2	57.6	14.6	88.3	46.2	59.5	37.3

Table 3. Experimental results of multipitch analysis based on the tatum-level model for 30 piano pieces labeled as `ENSTDkCl`.

Language model (prior distribution on $\boldsymbol{S}$)	Frame-level evaluation				Tatum-level evaluation			
	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{F}_{on}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{F}_{on}$
Uniform model	89.2	41.3	55.1	21.5	89.7	40.9	54.9	21.8
Sparse model ($\alpha = 12.5$)	77.4	52.5	60.7	40.3	78.3	52.4	60.9	41.9
Key-aware model ($\alpha = 12.5$)	73.8	68.6	69.6	53.9	74.9	68.6	70.1	55.2
Chord-aware model ($\alpha = 12.5$)	74.5	68.5	**70.0**	**55.6**	75.6	68.6	**70.5**	**57.0**
Chord-aware Markov model ($\alpha = 12.5$)	84.6	51.2	62.2	34.3	85.3	50.9	62.3	36.9
Sparse model ($\alpha = 1$)	88.3	44.1	57.5	24.0	88.9	43.7	57.3	24.3
Chord-aware model ($\alpha = 1$)	89.2	41.5	55.3	21.6	89.7	41.0	55.1	21.8

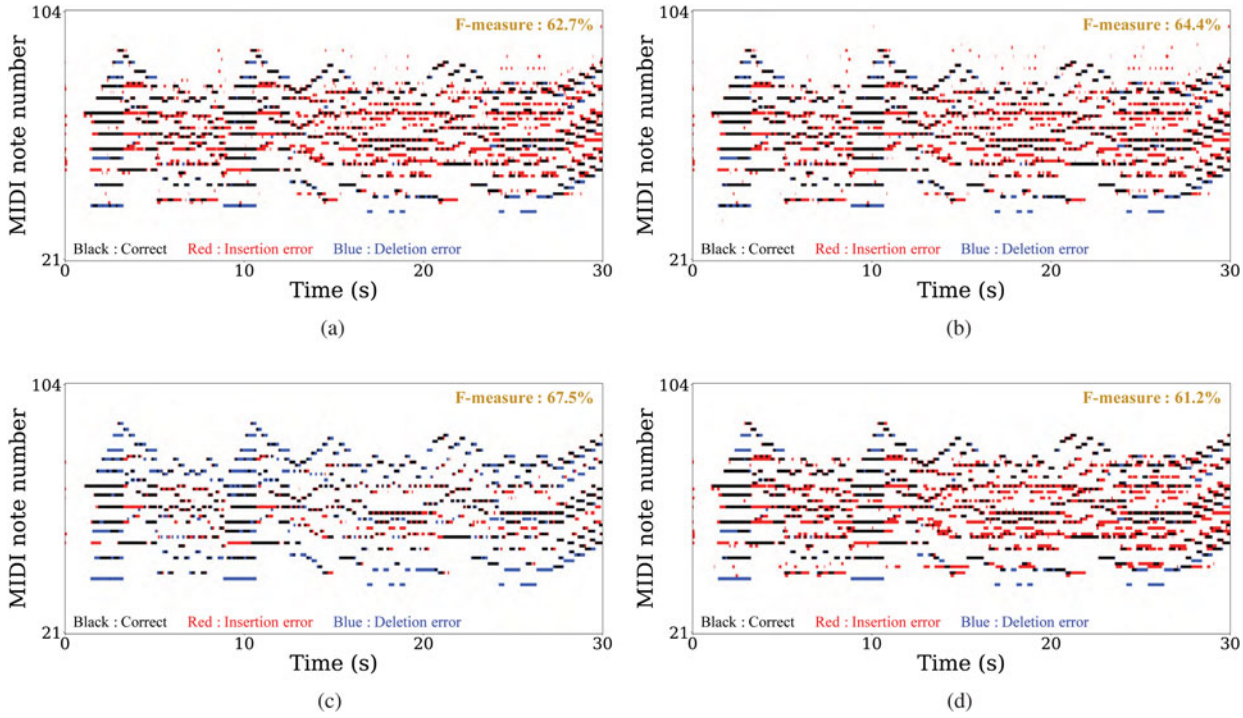

Fig. 7. Piano rolls estimated for MUS-chpn_p19_ENSTDkCl. (a) Sparse model, (b) chord-aware model ($\alpha = 1$), (c) chord-aware model ($\alpha = 12.5$), (d) chord-aware Markov model ($\alpha = 12.5$).

to separately deal with the dependencies of each musical note on the previous note and the current chord.

The similar results were obtained by the tatum-level model. Contrary to expectations, the *F*-measures obtained by the tatum-level model were not as good as those obtained by the frame-level model in the tatum-level evaluation. The piano rolls estimated by both models are shown in Fig. 8, where the tatum-level model outperformed and underperformed the frame-level model in the upper and lower examples, respectively. The tatum-level model tended to overestimate the durations of musical notes because the acoustic likelihood used for sampling a binary variable S_{kn} is given by the product of the acoustic likelihoods of all time frames contained in tatum unit n. If the likelihood of $S_{kn} = 1$ was larger by several orders of magnitude than that of $S_{kn} = 0$ at a time frame, the acoustic likelihood of the tatum unit tended to support $S_{kn} = 1$. The tatum-level model thus performed worse than the frame-level model, even in the tatum-level evaluation, when musical notes (e.g., triplet or arpeggio) that cannot be represented on a 16th-note-level grid were included in a target piece (Fig. 8). In such cases, the tatum-level model worked well, as shown in the upper example.

2) Evaluation of prior training

We evaluated the effectiveness of prior training of the language model via leave-one-out cross validation in which one musical piece was used for evaluation and the others were used for training the language model. We compared the performances of the following three conditions:

(a) *Baseline*: The whole model was trained in an unsupervised manner. This model is the same as the chord-aware model (d) in Section V-C-1.
(b) *Learning from piano rolls without chord annotations*: The ground-truth piano rolls were used for training the language model while estimating underlying chords.
(c) *Learning from piano rolls with chord annotations*: The ground-truth piano rolls with ground-truth chord annotations were used for training the language model.

As shown in Tables 4 and 5, the note-onset *F*-measures were improved by 1.8 pts in the frame-level model and 1.1 pts in the tatum-level model thanks to the improvement of the precision rate when the language model was trained by using piano rolls with chord annotations. We found that musically unnatural short musical notes can be avoided by using the pretrained language model, i.e., considering the note components of chords. On the other hand, the frame-wise *F*-measures remained almost the same, while the accuracy of chord estimation was improved by 15.0% in the frame-level model and 14.1% in the tatum-level model. One reason for this would be that the musical notes that were wrongly detected by the baseline method and were corrected by the language model were very short and had little impact on the frame-wise *F*-measures. Another reason is that the typical note cooccurrences could be learned as chords even in an unsupervised condition. We thus need to incorporate other musical structures (e.g., rhythm structures or phrase structures) in the language model.

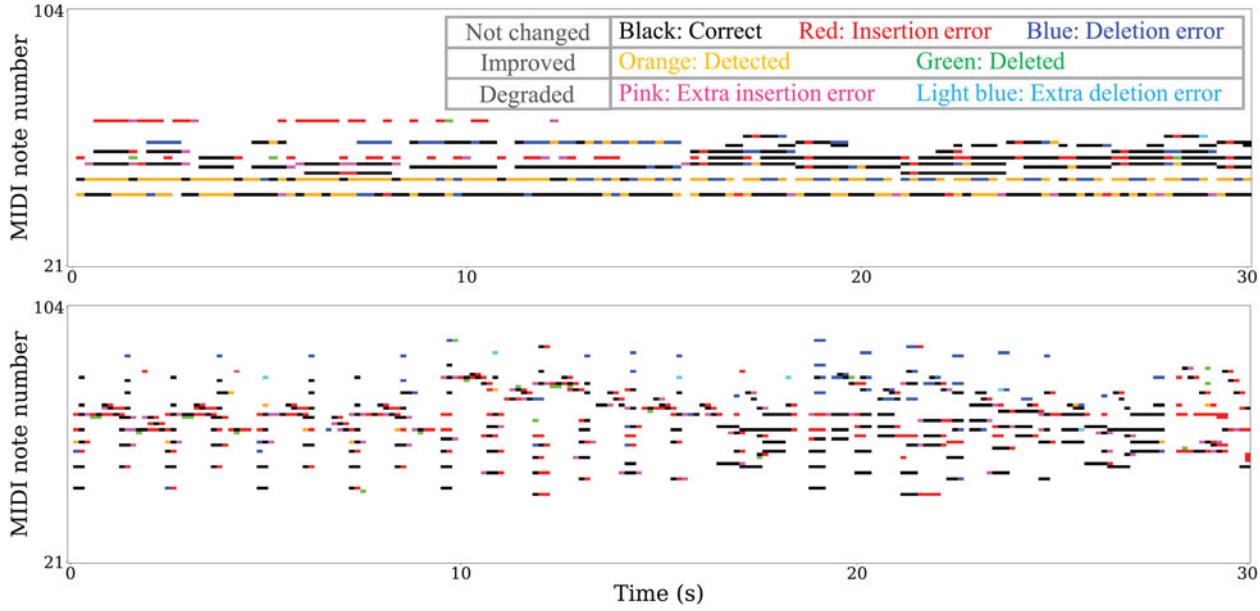

Fig. 8. Piano rolls of two musical pieces estimated by using the frame-level and tatum-level models.

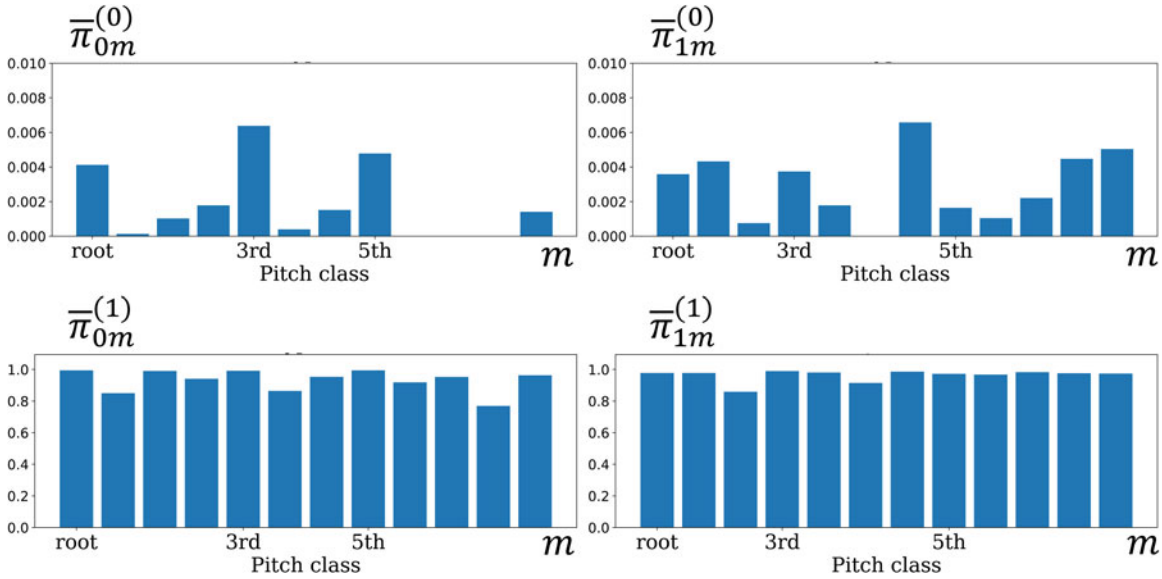

Fig. 9. The emission probabilities $\overline{\pi}$ estimated for MUS-chpn_p19_ENSTDkCl.

Table 4. Experimental results of multipitch analysis based on the pre-trained frame-level model.

Prior training using		Frame-level evaluation			
Piano rolls	Chords	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{F}_{on}$
		73.7	69.0	**69.8**	49.4
✓		72.3	70.4	69.6	50.5
✓	✓	71.6	71.0	69.5	**51.2**

Table 5. Experimental results of multipitch analysis based on the pre-trained tatum-level model.

Prior training using		Frame-level evaluation			
Piano rolls	Chords	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{F}_{on}$
		75.6	68.6	**70.5**	57.0
✓		75.1	69.2	70.5	57.2
✓	✓	74.3	69.9	70.3	**58.1**

3) Comparison with existing methods

We compared the performance of the proposed model with four existing unsupervised models: two PLCA-based models proposed by Benetos *et al.* [16] and by Berg-Kirkpatrick *et al.* [17] and two NMF-based models proposed by Vincent *et al.* [4] and by O'Hanlon *et al.* [20]. These models were trained using audio data produced by pianos that were not included in the test data.

Table 6 shows the performances reported in [17] and obtained by the proposed model. Our model, required no

Table 6. Performance comparison between five methods.

Method	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{F}_{on}$
Proposed model	73.7	69.0	69.8	49.4
Benetos *et al.* [16]	–	–	68.0	68.6
Berg-Kirkpatrick *et al.* [17]	80.7	69.1	**74.4**	**76.4**
Vincent *et al.* [4]	63.6	79.6	70.7	69.0
O'Hanlon *et al.* [20]	72.8	73.4	73.2	58.3

prior training, outperformed the NMF-based method [16] by 1.8 pts in terms of the frame-level *F*-measure. This considered to be promising because the NMF-based models [4, 20] are purely based on acoustic modeling and could be extended in the same way as the proposed model.

VI. CONCLUSION

This paper presented a unified statistical model for multipitch analysis that can jointly estimate pitches and chords from music signals in an unsupervised manner. The proposed model consists of an acoustic model (Bayesian NMF) and a language model (Bayesian HMM), and both models can contribute to estimating a piano roll. When a piano roll is given, these models can be updated independently. The piano roll can then be estimated considering the difference in time resolution between the two models.

The experimental results showed the potential of the proposed method for unified music transcription and grammar induction. Although the performance of multipitch estimation was improved by iteratively updating the language and acoustic models, the proposed model did not always outperform other existing methods. The main reason is that simplified acoustic and language models (shift invariance of basis spectra and local dependency of pitches) are used in the current model because the main goal of this paper is to show the effectiveness of integrating the acoustic and language models and the feasibility of unsupervised joint estimation of chords and pitches.

We plan to integrate the state-of-the-art acoustic models such as [20] and [15] with our language model. To improve the language model, we need to deal with music grammar of chords, rhythms, and keys. Probabilistic rhythm models, for example, have already been proposed by Nakamura *et al.* [39], which could be integrated with our language model. Moreover, we try to use a deep generative model as a language model to learn more complicated music grammar. Since the performance of the proposed method is considered to be degraded if we use tatum times obtained by a beat tracking method instead of using correct tatum times, joint estimation of tatum times, pitches, and chords is another important direction of research.

Our approach has a deep connection to language acquisition. In the field of natural language processing, unsupervised grammar induction from a sequence of words and unsupervised word segmentation for a sequence of characters have actively been studied [40, 41]. Since our model can directly infer music grammars (e.g., chord compositions) from either musical scores (discrete symbols) or music signals, the proposed technique is expected to be useful for the emerging topic of language acquisition from continuous speech signals [42].

FINANCIAL SUPPORT

This study was partially supported by JSPS KAKENHI No. 26700020 and No. 16H01744, JSPS Grant-in-Aid for Fellows No. 16J05486, and JST ACCEL No. JPMJAC1602.

REFERENCES

[1] Smaragdis, P.; Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, 177–180.

[2] Hoffman, M.; Blei, D.M.; Cook, P.R.: Bayesian nonparametric matrix factorization for recorded music, in *Int. Conf. on Machine Learning (ICML)*, 2010, 439–446.

[3] Virtanen, T.; Klapuri, A.: Analysis of polyphonic audio using source-filter model and non-negative matrix factorization, in *NIPS Workshop on Advances in Models for Acoustic Processing*, 2006.

[4] Vincent, E.; Bertin, N.; Badeau, R.: Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech, Language Process.*, **18** (3) (2010), 528–537.

[5] Rocher, T.; Robine, M.; Hanna, P.; Strandh, R.: Dynamic chord analysis for symbolic music, in *Int. Computer Music Conf. (ICMC)*, 2009, 41–48.

[6] Sheh, A.; Ellis, D.P.: Chord segmentation and recognition using EM-trained hidden Markov models, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2003, 185–191.

[7] Maruo, S.; Yoshii, K.; Itoyama, K.; Mauch, M.; Goto, M.: A feedback framework for improved chord recognition based on NMF-based approximate note transcription, in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, 196–200.

[8] Ueda, Y.; Uchiyama, Y.; Nishimoto, T.; Ono, N.; Sagayama, S.: HMM-based approach for automatic chord detection using refined acoustic features, in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, 5518–5521.

[9] Liang, D.; Hoffman, M.: Beta process non-negative matrix factorization with stochastic structured mean-field variational inference, in *NIPS Workshop on Advances in Variational Inference*, 2014.

[10] Ojima, Y.; Nakamura, E.; Itoyama, I.; Yoshii, K.: A hierarchical Bayesian model of chords, pitches, and spectrograms for multipitch analysis, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2016, 309–315.

[11] Emiya, V.; Badeau, R.; David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Trans. Audio, Speech, Language Process.*, **18** (6) (2010), 1643–1654.

[12] Ycart, A.; Benetos, E.: A-MAPS: Augmented MAPS dataset with rhythm and key annotations, in *Int. Society for Music Information Retrieval Conf. (ISMIR), Late Breaking Demo*, 2018.

[13] Cemgil, A.T.: Bayesian inference for nonnegative matrix factorisation models. *Comput. Intell. Neurosci.*, **2009** (ID:785152) (2009, 1–17.

[14] Durrieu, J.L.; Richard, G.; David, B.; Févotte, C.: Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Trans. Audio, Speech, Language Process.*, **18** (3) (2010), 564–575.

[15] Cheng, T.; Mauch, M.; Benetos, E.; Dixon, S.: An attack/decay model for piano transcription, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2016, 584–590.

[16] Benetos, E.; Weyde, T.: Explicit duration hidden Markov models for multiple-instrument polyphonic music transcription, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2013, 269–274.

[17] Berg-Kirkpatrick, T.; Andreas, J.; Klein, D.: Unsupervised transcription of piano music, in *Advances in Neural Information Processing Systems (NIPS)*, 2014, 1538–1546.

[18] Benetos, E.; Weyde, T.: An efficient temporally-constrained probabilistic model for multiple-instrument music transcription, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2015, 701–707.

[19] Smaragdis, P.: Convolutive speech bases and their application to speech separation. *IEEE Trans. Audio, Speech, Language Process.*, **15** (1) (2007), 1–14.

[20] O'Hanlon, K.; Plumbley, M.D.: Polyphonic piano transcription using non-negative matrix factorisation with group sparsity, in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, 3112–3116.

[21] Nam, J.; Ngiam, J.; Lee, H.; Slaney, M.: A classification-based polyphonic piano transcription approach using learned feature representations, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2011, 175–180.

[22] Boulanger-Lewandowski, N.; Bengio, Y.; Vincent, P.: High-dimensional sequence transduction, in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, 3178–3182.

[23] Hamanaka, M.; Hirata, K.; Tojo, S.: Implementing "A Generative Theory of Tonal Music". *J. New. Music. Res.*, **35** (4) (2006), 249–277.

[24] Jackendoff, R.; Lerdahl, F.: A Generative Theory of Tonal Music. MIT Press, Cambridge, Massachusetts, 1985.

[25] Nakamura, E.; Hamanaka, M.; Hirata, K.; Yoshii, K.: Tree-structured probabilistic model of monophonic written music based on the generative theory of tonal music, in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, 276–280.

[26] Hu, D.; Saul, L.K.: A probabilistic topic model for unsupervised learning of musical key-profiles, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2009, 441–446.

[27] Raczyński, S.; Vincent, E.; Bimbot, F.; Sagayama, S.: Multiple pitch transcription using DBN-based musicological models, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2010, 363–368.

[28] Raczyński, S.; Vincent, E.; Sagayama, S.: Dynamic Bayesian networks for symbolic polyphonic pitch modeling. *IEEE Trans. Audio, Speech, Language Process.*, **21** (9) (2013), 1830–1840.

[29] Böck, S.; Schedl, M.: Polyphonic piano note transcription with recurrent neural networks, in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, 121–124.

[30] Sigtia, S.; Benetos, E.; Dixon, S.: An end-to-end neural network for polyphonic piano music transcription. *IEEE Trans. Audio, Speech, Language Process.*, **24** (5) (2016), 927–939.

[31] Holzapfel, A.; Benetos, E.: The Sousta corpus: Beat-informed automatic transcription of traditional dance tunes, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2016, 531–537.

[32] Ycart, A.; Benetos, E.: A study on LSTM networks for polyphonic music sequence modelling, in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2017, 421–427.

[33] Smaragdis, P.; Raj, B.; Shashanka, M.: Sparse and shift-invariant feature extraction from non-negative data, in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, 2069–2072.

[34] Cemgil, A.T.; Dikmen, O.: Conjugate gamma Markov random fields for modelling nonstationary sources, in *Independent Component Analysis and Signal Separation*, 2007, 697–705.

[35] Benetos, E.; Dixon, S.: Multiple-instrument polyphonic music transcription using a convolutive probabilistic model, in *Sound and Music Computing Conf. (SMC)*, 2011, 19–24.

[36] Schörkhuber, C.; Klapuri, A.; Holighaus, N.; Dörfler, M.: A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution, in *Audio Engineering Society Conf.*, 2014, 1–8.

[37] Fitzgerald, D.: Harmonic/percussive separation using median filtering, in *Int. Conf. on Digital Audio Effects (DAFx)*, 2010, 1–4.

[38] Dixon, S.: On the computer recognition of solo piano music, in *Australasian Computer Music Conf.*, 2000, 31–37.

[39] Nakamura, E.; Yoshii, K.; Sagayama, S.: Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices. *IEEE Trans. Audio, Speech, Language Process.*, **25** (4) (2017), 794–806.

[40] Johnson, M.: Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure, in *Annual Meeting of the Association of Computational Linguistics (ACL)*, 2008, 398–406.

[41] Mochihashi, D.; Yamada, T.; Ueda, N.: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling, in *Annual Meeting of the Association of Computational Linguistics (ACL)*, 2009, 100–108.

[42] Taniguchi, T.; Nagasaka, S.; Nakashima, R.: Nonparametric Bayesian double articulation analyzer for direct language acquisition from continuous speech signals. *IEEE Trans. Cogn. Develop. Syst.*, **8** (3) (2016), 171–185.

Yuta Ojima received the M.S. degree in informatics from Kyoto University, Kyoto, Japan, in 2018. His expertise is automatic music transcription.

Eita Nakamura received the Ph.D. degree in physics from the University of Tokyo, Tokyo, Japan, in 2012. After having been a Postdoctoral Researcher at the National Institute of Informatics, Meiji University, and Kyoto University, Kyoto, Japan, he is currently a Research Fellow of Japan Society for the Promotion of Science. His research interests include music modeling and analysis, music information processing, and statistical machine learning.

Katsutoshi Itoyama received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2008 and 2011, respectively. He had been an Assistant Professor at the Graduate School of Informatics, Kyoto University, until 2018 and is currently a Senior Lecturer at Tokyo Institute of Technology. His research interests include musical sound source separation, music listening interfaces, and music information retrieval.

Kazuyoshi Yoshii received the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2005 and 2008, respectively. He had been a Senior Lecturer and is currently an Associate Professor at the Graduate School of Informatics, Kyoto University, and concurrently the Leader of the Sound Scene Understanding Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. His research interests include music analysis, audio signal processing, and machine learning.