Bayesian Singing Transcription Based on a Hierarchical Generative Model of Keys, Musical Notes, and F0 Trajectories

Ryo Nishikimi, Eita Nakamura, Masataka Goto, Katsutoshi Itoyama, *Member, IEEE* and Kazuyoshi Yoshii *Member, IEEE*

***Abstract*— This paper describes automatic singing transcription (AST) that estimates a human-readable musical score of a sung melody represented with quantized pitches and durations from a given music audio signal. To achieve the goal, we propose a statistical method for estimating the musical score by quantizing a trajectory of vocal fundamental frequencies (F0s) in the time and frequency directions. Since vocal F0 trajectories considerably deviate from the pitches and onset times of musical notes specified in musical scores, the local keys and rhythms of musical notes should be taken into account. In this paper we propose a Bayesian hierarchical hidden semi-Markov model (HHSMM) that integrates a musical score model describing the local keys and rhythms of musical notes with an F0 trajectory model describing the temporal and frequency deviations of an F0 trajectory. Given an F0 trajectory, a sequence of musical notes, that of local keys, and the temporal and frequency deviations can be estimated jointly by using a Markov chain Monte Carlo (MCMC) method. We investigated the effect of each component of the proposed model and showed that the musical score model improves the performance of AST.**

***Index Terms*—Automatic singing transcription, hierarchical hidden semi-Markov model**

I. Introduction

Automatic singing transcription (AST) refers to estimating a musical score of a sung melody from a music audio signal (Fig. 1). It forms the basis of music information retrieval (MIR) because a melody is the most salient part of popular music that affects the impression of a musical piece. Generally, a sung melody is represented as a sequence of musical notes with pitches quantized in semitones and onset and offset times quantized in certain time units (tatums). To realize AST, one needs to quantize continuous physical quantities such as vocal fundamental frequencies (F0s) in both time and frequency directions. Since a lot of effort has been devoted to estimating vocal F0 trajectories from music audio signals (see Section II), we focus on AST from vocal F0 trajectories.

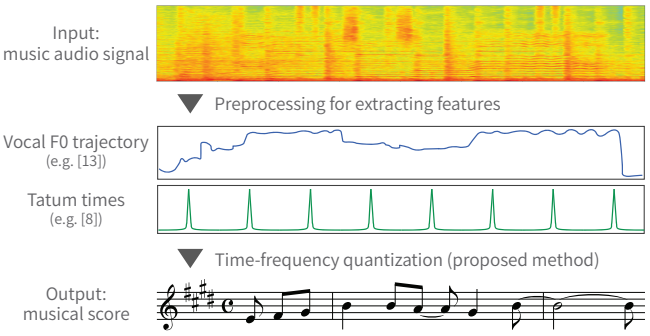


Fig. 1. The problem of automatic singing transcription. The proposed method takes as input a vocal F0 trajectory and tatum times, and estimates a sequence of musical notes by quantizing the F0 trajectory in the time and frequency directions.

One of the major difficulties of AST is that continuous F0 trajectories include temporal and frequency deviations from straight pitch trajectories indicated in scores. This prohibits a simple quantization method (called *majority-vote method*) that estimates a pitch as the majority of F0s in each tatum interval. A promising way to obtain a natural score is to integrate a *musical score model* that describes the organization of notes in scores with an *F0 trajectory model* representing the temporal and frequency deviations. This framework is similar to the statistical speech recognition approach based on a *language model* and an *acoustic model* [1]. Recent studies have applied musical score models for music transcription in the framework of probabilistic modeling [2], [3] and deep learning [4], [5].

To build a musical score model, we focus on how pitches and rhythms of musical notes are structured in a sung melody. In tonal music, pitches have sequential interdependence and are controlled by underlying musical keys or scales. Onset times in scores also have sequential interdependence and are controlled by underlying metrical structure. To represent such characteristics, it is necessary to formulate a musical score model in the musical-note level, instead of in time-frame level [4], [5]. On the other hand, a vocal F0 trajectory is represented in the time-frame level, or possibly in the tatum level after applying beat tracking. Because of the mismatch of time scales, integration of a note-level musical score model and a frame- or tatum-level F0/acoustic model poses a challenge in

Manuscript received XXXX XX, 2019; revised XXXX XX, 2019; accepted XXXX XX, 2019. Date of publication XXXX XX, 2019; date of current version XXXX XX, 2019. This work was partially supported by JSPS KAKENHI Grant Numbers No. 19H04137, No. 19K20340, No. 16H01744, No. 16J05486, No. 19J15255, and No. 19K12017, JST ACCEL No. JPMJAC1602, and Kayamoni Foundation. The associate editor coordinating the review of this manuscript and approving it for publication is Prof. XXX YYY (Corresponding author: Ryo Nishikimi).

R. Nishikimi, E. Nakamura, and K. Yoshii are with Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan, email: {nishikimi, enakamura, yoshii}@sap.ist.i.kyoto-u.ac.jp. E. Nakamura is also with the Hakubi Center for Advanced Research, Kyoto University, Kyoto 606-8501, Japan.

K. Itoyama is with School of Engineering, Tokyo Institute of Technology, Tokyo 152-8552, Japan, email: itoyama@ra.sc.e.titech.ac.jp.

M. Goto is with National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki 305-8568, Japan, email: m.goto@aist.go.jp.

Digital Object Identifier XXXX

probabilistic modeling, which is still open [6].

For key- and rhythm-aware AST, we previously proposed a hierarchical hidden semi-Markov model (HHSMM) [7] that consists of a musical score model and an F0 trajectory model under an condition that the tatum times are given in advance or estimated by a beat detection method [8] (Fig. 2). The musical score model generates a note sequence and consists of three sub-models describing local keys, pitches, and onset score times (Section III-B). The local keys are sequentially generated by a Markov model and the pitches of musical notes are then sequentially generated by another Markov model conditioned on the local keys. The onset score times are sequentially generated by a metrical Markov model [9], [10] defined on the tatum grid. The F0 trajectory model describes the temporal and frequency deviations added to a step-function-like pitch trajectory corresponding to the generated score (Section III-C). To stably learn the musical characteristics unique to each musical piece from a small amount of piece-specific data, the HHSMM is formulated in a Bayesian manner (Section III-D).

To estimate a latent sequence of musical notes with decent durations from an observed vocal F0 trajectory by using the HHSMM, in this paper we propose a combination of an iterative Gibbs sampler and a modified Viterbi algorithm that is penalized for intensely favoring longer notes with less frequent transitions (Section IV-C). The whole model can be estimated in an unsupervised or semi-supervised manner (Sections IV-A and IV-B) by optimizing on the fly or pretraining the musical score model, respectively. Since putting more emphasis on the musical score model was shown to be effective in our previous work [7], in this paper we carefully optimize the weighting factors on the individual components of the musical score and F0 trajectory models and the note duration penalization with Bayesian optimization [11] or grid search (Section V-A2).

The main contributions of this study are as follows. First, we provide a full description of the HHSMM (Section III) that is used for transcribing a human-readable score consisting of quantized pitches and onset times from a music audio signal (monophonic F0 trajectory) via the improved learning methods (Section IV). This is a principled statistical approach to a well-known open problem of how to integrate a note-level language model with a tatum- or frame-level acoustic model in automatic music transcription. Second, we found that the rhythm and key models of the musical score model and the note duration penalization were particularly effective, by conducting comprehensive comparative experiments for investigating the performances of the unsupervised and semi-supervised learning methods (Section V-B1) and evaluating the musical score model (Section V-B2), the F0 trajectory model (Section V-B3), and the note duration penalization (Section V-B4).

Section II introduces related work on AST. Sections III and IV describe our statistical approach to AST (generative modeling and posterior inference). Section V reports the results of comparative experiments. Section VI summarizes the paper.

II. Related Work

A typical approach to AST consists of three steps: melody extraction, piano-roll estimation, and rhythm transcription. The melody extraction refers to estimating predominant F0 trajectories from music audio signals. The piano-roll estimation refers to converting the vocal F0s to a piano-roll representation, *i.e.*, a sequence of note events represented by onset times, durations, and quantized pitches. While the pitches take discrete values, the onset times and durations take continuous values. To estimate a complete musical score, *i.e.*, a sequence of musical notes, the rhythm transcription is performed to quantize onset and offset times in units of tatums. Given an estimated F0 trajectory with tatum positions, the proposed method simultaneously performs the piano-roll estimation and rhythm transcription by modeling the generative process of a vocal F0 trajectory.

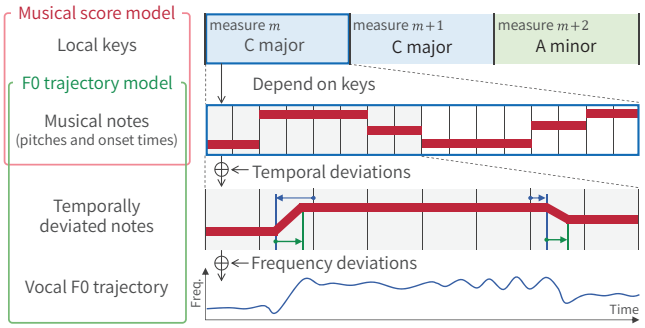


Fig. 2. The generative process of a vocal F0 trajectory based on the proposed model consisting of a musical score model and an F0 trajectory model. The musical score model represents the generative process of musical notes (pitches and onset score times) based on local keys assigned to measures. In the figure of musical notes, the black vertical lines represent a tatum grid given as input. The F0 trajectory model represents the generative process of a vocal F0 trajectory from the musical notes by adding the frequency and temporal deviations. In the figure of temporally deviated notes, the arrows represent temporal deviations of onset times from tatum times.

A. Melody Extraction

Estimation of vocal F0 trajectories for music audio signals, *a.k.a.* melody extraction, has been actively studied [12]–[17]. One of the most basic methods is subharmonic summation (SHS) [12], which calculates the sum of the harmonic components of each candidate F0. Ikemiya *et al.* [13] improved F0 estimation based on the combination of SHS and robust principle component analysis (RPCA) [18]. Salamon *et al.* [14] estimated several F0 contours according to a salience function and then recursively removed non-vocal contours by focusing on the singing characteristics. Durrieu *et al.* [15] extracted the main melody by representing the melody part with a source-filter model and accompaniments with a model inspired by non-negative matrix factorization (NMF). Mauch *et al.* [17] modified the YIN [16] in a probabilistic manner so that it would determine multiple candidate F0s.

B. Piano-roll Estimation

Piano-roll estimation has also been studied [19]–[26]. A straightforward quantization method that takes the majority of vocal F0s in each tatum interval is often used [19]. Paiva *et al.* [20] proposed a cascading method that performs multipitch estimation, pitch trajectory construction and segmentation, and

extraction of musical notes that form a main melody. Laaksonen *et al.* [22] divided audio signals into segments corresponding to note events by using the boundaries of chords as input, and then estimated the pitch of each note event by using a scoring function based on chord and key information. Ryynänen *et al.* [23] proposed a hierarchical HMM (HHMM) that categorizes the F0s of each note region into attack, sustain, and release states. This model represents the pitch transitions in an upper-level Markov chain and the transitions between the vocal fluctuation types in a lower-level Markov chain. Molina *et al.* [24] focused on the hysteresis characteristics of vocal F0s. Yang *et al.* [25] proposed an HHMM that represents the generative process of an F0 trajectory in a two-dimensional phase plane spanned by F0s and their first derivations (ΔF0s). Mauch *et al.* [26] developed a software called Tony that estimates a vocal F0 trajectory from a music audio signal using probabilistic YIN (pYIN) [17] and then estimates note events using a modified version of Ryynänen's method [23].

C. Rhythm Transcription

The goal of rhythm transcription is to quantize onset and offset times of MIDI-like note events for appropriate score representation [9], [10], [27]–[30]. A major approach to this problem is to formulate an HMM [9], [10], [27]. Takeda *et al.* [27], for example, proposed a duration-based HMM that represents note values as latent variables and actual note durations as observed variables. Other studies have proposed onset-based HMMs called metrical HMMs that represent the note onset positions on the tatum grids as latent variables and the actual onset times as observed variables [9], [10]. The rhythm and onset deviation models used by this method are similar to those used by our method. Note that both types of HMM-based methods can be extended for polyphonic rhythm transcription (*e.g.*, for MIDI piano performances). A probabilistic context-free grammar (PCFG) [28] and a connectionist approach [29] have also been studied for rhythm transcription. It has been shown that HMMs are currently state of the art [30].

III. Generative Modeling

This section defines the task of AST (Section III-A) and explains the hierarchical hidden semi-Markov model (HHSMM) that consists of a musical score model and an F0 trajectory model (Fig. 2). The musical score model represents the generative process of sung notes in the tatum level (Section III-B) and the F0 trajectory model represents the generative process of vocal F0s in the frame level from the note sequence (Section III-C). We introduce prior distributions to complete Bayesian formulation. This is effective for estimating the reasonable parameters of the proposed model from a small amount of data (Section III-D). We define the meanings of several terms regarding temporal information as follows:

- Onset/offset times and duration: the start/end times and length of a note represented in the frame level.
- Onset/offset score times and note value: the start/end times and length of a note represented in the tatum level.
- Tatum position: the relative position of a tatum in a measure including the tatum.

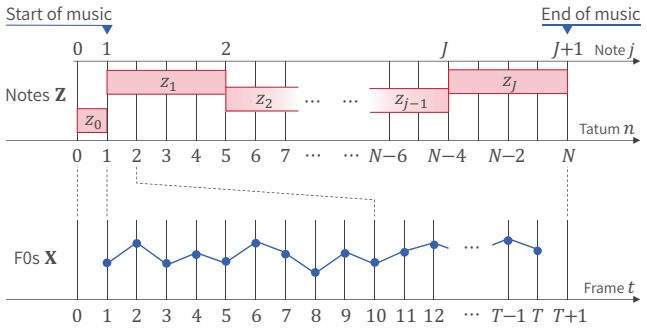


Fig. 3. Relationships between different time indices j, n, and t. The upper figure shows the start and end beat times of each musical note indexed by j. The dotted lines between the upper and lower figures represent correspondence between the tatum index n and the frame index t. The lower figure shows the F0 value of each time frame. The onset of the first note z_1 is the start of music and z_0 is a supplementary note that is used only for calculating the slanted line representing the transient segment of z_1.

A. Problem Specification

Our problem is formalized as follows (Figs. 1 and 3):

Input:
A frame-level vocal F0 trajectory $\mathbf{X} = x_{0:T}$
and tatum times $\mathbf{Y} = y_{0:N} = (t_n, l_n)_{0:N}$
Output:
A sequence of musical notes $\mathbf{Z} = z_{0:J} = (p_j, o_j)_{0:J}$
By-product:
A sequence of local keys $\mathbf{S} = s_{0:M}$

where $x_{0:T} = \{x_0, \ldots, x_T\}$ etc., and T, N, J, and M indicate the number of frames, tatums, estimated notes, and measures, respectively. The time-shifting interval is 10 ms in this study. x_t indicates a log F0 in cents at frame t, where unvoiced frames are represented as $x_t = \mathrm{uv}$. t_n indicates a frame corresponding to tatum n, where $t_0 = 0$, $t_1 = 1$, and $t_N = T+1$. $l_n \in \{1, \ldots, L\}$ indicates the tatum position, where L is the number of tatums included in a measure ($L = 16$ in this paper) and $l_n = 1$ indicates the barline. Each note z_j is represented as a pair of a semitone-level pitch $p_j \in \{1, \ldots, K\}$ and an onset score time $o_j \in \{0, \ldots, N\}$, where K is the number of unique pitches considered (*e.g.*, $K = 88$ pitches from A0 to C8), $o_0 = 0$, $o_1 = 1$, $o_{J+1} = N$. We introduce local keys $s_{0:M}$ for each measure. The local key s_m of measure m takes a value in $\{\mathrm{C}, \mathrm{C\#}, \cdots, \mathrm{B}\} \times \{\text{major, minor}\}$ (the tonic is represented as $\mathrm{C}{=}0, \mathrm{C\#}{=}1, \ldots, \mathrm{B}{=}11$, and the local keys are numbered from 1 to 24). We have introduced supplementary variables x_0, y_0, z_0, and s_0 in order to ease the handling of latent variables at the beginning of music.

In this paper, we deal with songs in the pop music style. It is assumed that a target piece is in $4/4$ and that tatum unit is 16th note. Rests, notes shorter than the tatum unit, and triplets are not considered. Offset score times are not explicitly modeled, *i.e.*, the offset score time of each note corresponds to the onset score time of the next note. It is also assumed that the maximum distance between successive onset score time (*i.e.*, maximum note value) is L.

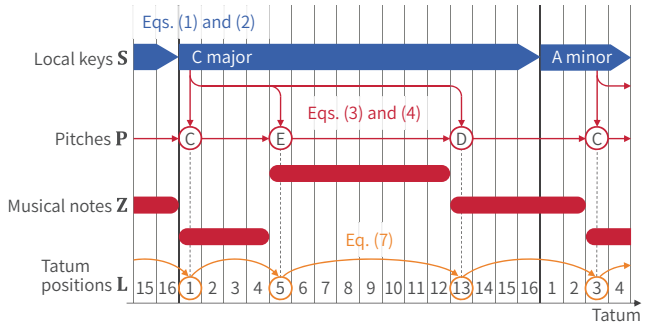


Fig. 4. A musical score model that represents the generative process of local keys, note pitches, and note onsets. The top row represents the Markov chain of local keys. The second row represents the Markov chain of the note pitches. The third row represents a sequence of musical notes. The bottom represents the Markov chain of the onset score times of musical notes. The vertical lines represent tatum times and the bold ones represent bar lines.

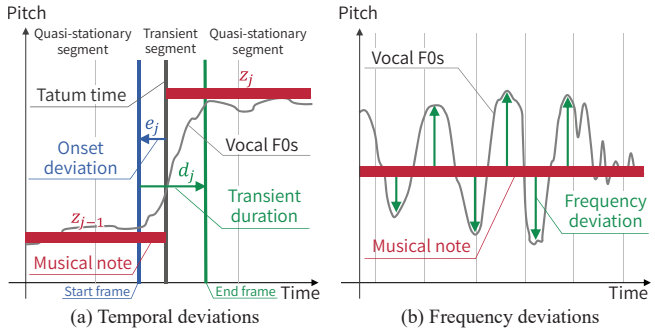


Fig. 5. Temporal and frequency deviations in vocal F0 trajectories. In both figures, the black vertical lines represent tatum times. In Fig. (a), the blue and green vertical lines represent the start and end frames of the transient segment of a vocal F0 trajectory. In Fig. (b), the arrows represent the frequency deviations of a vocal F0 trajectory from the frequency of a musical note.

B. Musical Score Model

The musical score model represents the generative process of local keys $\mathbf{S}$ and musical notes $\mathbf{Z} = \{\mathbf{P}, \mathbf{O}\}$. More specifically, local keys $\mathbf{S}$ are sequentially generated by a Markov model and pitches $\mathbf{P}$ are then sequentially generated by another Markov model conditioned on $\mathbf{S}$ (Fig. 4). With an independent process, onset score times $\mathbf{O}$ are sequentially generated by a metrical Markov model [9], [10]. We henceforth omit to explicitly write the dependency on $\mathbf{Y}$ for brevity.

1) Model for Local Keys: To consider the relevance of adjacent local keys (*e.g.*, the local keys are likely to change infrequently), the local keys $\mathbf{S}$ are assumed to follow a first-order Markov model as follows:

$$s_0 \sim \text{Categorical}(\boldsymbol{\pi}_0), \tag{1}$$

$$s_m \mid s_{m-1} \sim \text{Categorical}(\boldsymbol{\pi}_{s_{m-1}}), \tag{2}$$

where $\boldsymbol{\pi}_0 \in \mathbb{R}_+^{24}$ and $\boldsymbol{\pi}_s \in \mathbb{R}_+^{24}$ are initial and transition probabilities. We write $\boldsymbol{\pi} = \boldsymbol{\pi}_{0:24}$. Given the similarities between keys (*e.g.*, relative transitions from C major would be similar to those from D major), a hierarchical Dirichlet or Pitman-Yor language model [31] with a shared prior generating key-specific priors and distributions would be useful for precise key modeling.

2) Model for Pitches: The pitches $\mathbf{P}$ are assumed to follow a first-order Markov model conditioned on the local keys $\mathbf{S}$ as follows:

$$p_0 \mid \mathbf{S} \sim \text{Categorical}(\boldsymbol{\phi}_{s_0,0}), \tag{3}$$

$$p_j \mid p_{j-1}, \mathbf{S} \sim \text{Categorical}(\boldsymbol{\phi}_{s_{m_j},p_{j-1}}), \tag{4}$$

where $\boldsymbol{\phi}_{s0} \in \mathbb{R}_+^K$ and $\boldsymbol{\phi}_{sp} \in \mathbb{R}_+^K$ are initial and transition probabilities for pitches in local key s, and m_j denotes a measure to which the onset of note j belongs. Let $\boldsymbol{\phi} = \boldsymbol{\phi}_{1:24,0:K}$. We assume that the initial and transition probabilities in different local keys are related by a circular shift (change of tonic), and $\boldsymbol{\phi}$ are represented as follows:

$$\phi_{s0p'} \propto \bar{\phi}_{\text{type}(s),0,\deg(s,p')}, \tag{5}$$

$$\phi_{spp'} \propto \bar{\phi}_{\text{type}(s),\deg(s,p),\deg(s,p')}, \tag{6}$$

where $\text{type}(s) \in \{\text{major, minor}\}$ indicates the type of key s, $\deg(s,p) \in \{1,\ldots,12\}$ indicates the degree of pitch p in key s (if the tonic of key s is $[s]$, $\deg(s,p) = (p - [s] \mod 12) + 1$), and $\bar{\boldsymbol{\phi}}_{r0} \in \mathbb{R}_+^{12}$ and $\bar{\boldsymbol{\phi}}_{rh} \in \mathbb{R}_+^{12}$ indicate initial and transition probabilities under a local key of type $r \in \{\text{major, minor}\}$ and tonic C, where $h \in \{1,\ldots,12\}$ is a subscript representing a pitch degree. The proposed method learns only the probabilities of *relative* pitch degrees $\bar{\boldsymbol{\phi}}$ in an unsupervised or semi-supervised manner. The probabilities of *absolute* pitches $\boldsymbol{\phi}$ are then obtained by expanding $\bar{\boldsymbol{\phi}}$ according to Eqs. (5) and (6) and used for estimating a sequence of musical notes. In other words, the same transition probabilities of pitch degrees are used for every octave range, and for pitch transitions beyond an octave we use the probabilities of the corresponding pitch transitions within an octave with the same pitch degrees.

3) Model for Rhythms: The onset score times $\mathbf{O}$ are assumed to follow a metrical Markov model [9], [10] as follows:

$$l_{o_j} \mid l_{o_{j-1}} \sim \text{Categorical}(\boldsymbol{\lambda}_{l_{o_j-1}}), \tag{7}$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{1:L} \in \mathbb{R}_+^{L \times L}$ denotes transition probabilities for tatum positions, *i.e.*, $\lambda_{l,l'}$ $(l, l' \in \{1,\ldots,L\})$ indicates the transition probability from tatum position l to l'. We interpret that if $l_{o_{j-1}} < l_{o_j}$ the onsets of notes $j-1$ and j are in the same measure and if $l_{o_{j-1}} \geq l_{o_j}$ they are in the adjacent measures.

C. F0 Trajectory Model

The F0 trajectory model represents the generative process of an F0 trajectory $\mathbf{X}$ from a note sequence $\mathbf{Z}$. We consider both temporal and frequency deviations (Fig. 5).

1) Model for Temporal Deviations: As shown in Fig. 5-(a), vocal F0s corresponding to each note are assumed to have a transient segment (*e.g.*, portamento) and a quasi-stationary segment (*e.g.*, vibrato). The actual onset time of note j, which is defined as the first frame of the transient segment, can deviate from the tatum time t_{o_j}. Let $e_j \in [e_{\min}, e_{\max}]$ be the deviation of the actual onset time from t_{o_j}, where $[e_{\min}, e_{\max}]$ indicates its range. The onset and offset time deviations at the start and end of the musical piece are fixed to zero ($e_1 = e_{J+1} = 0$), and the onset time deviation of the supplementary note z_0 is also set to zero for convenience ($e_0 = 0$). If $e_j < 0$ ($e_j > 0$), note j begins earlier (later) than t_{o_j}. Because $\mathbf{E} = e_{0:J}$ are

considered to be distributed according to a possibly multimodal distribution, in this paper we use a categorical distribution as the most basic distribution of discrete variables as follows:

$$e_j \sim \mathrm{Categorical}(\boldsymbol{\epsilon}), \tag{8}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}_+^{e_{\max}-e_{\min}+1}$ is a set of deviation probabilities.

Let $d_j \in \{1, \ldots, d_{\max}\}$ be the duration of the transient segment of note z_j, where $d_{\max}$ is the maximum number, and we set $d_0 = d_{J+1} = 1$. For the same reason as that for $\mathbf{E}$, we use a categorical distribution for $\mathbf{D} = d_{0:J}$ as follows:

$$d_j \sim \mathrm{Categorical}(\boldsymbol{\delta}), \tag{9}$$

where $\boldsymbol{\delta} \in \mathbb{R}_+^{d_{\max}}$ indicates a set of duration probabilities.

2) Model for Frequency Deviations: As shown in Fig. 5-(b), the vocal F0 trajectory $\mathbf{X}$ is generated by imparting frequency deviations to a temporally deviated pitch trajectory determined by the musical notes $\mathbf{Z}$, the onset time deviations $\mathbf{E}$, and the transient durations $\mathbf{D}$. Since vocal F0s can significantly deviate from score-indicated pitches, $\mathbf{X}$ are assumed to follow Cauchy distributions, which are more robust to outliers than Gaussian distributions, as follows:

$$x_t \mid \mathbf{Z}, \mathbf{E}, \mathbf{D} \sim \mathrm{Cauchy}(\mu_t, \sigma), \tag{10}$$

where μ_t and σ are the location and scale parameters, respectively. Note that if $x_t = \mathrm{uv}$, x_t is treated as missing data. The related studies [6], [7] also used the Cauchy distribution for frequency deviations as a better choice than the Gaussian distribution. As shown in Fig. 6, the actual duration of note j is given by $[t_{o_j}+e_j, t_{o_{j+1}}+e_{j+1})$ and the reference F0 trajectory is modeled as a slanted line in the transient segment and a horizontal line in the quasi-stable segment as follows (Fig. 6):

$$\mu_t = \begin{cases} [p_{j-1}] + \frac{([p_j]-[p_{j-1}])(t-(t_{o_j}+e_j))}{d_j} \\ \qquad (t \in [t_{o_j}+e_j, t_{o_j}+e_j+d_j)), \\ [p_j] \quad (t \in [t_{o_j}+e_j+d_j, t_{o_{j+1}}+e_{j+1})), \end{cases} \tag{11}$$

where $[p_j]$ indicates a log F0 [cents] corresponding to a semitone-level pitch p_j. Although F0 transitions between different pitches have complicated dynamics in reality, in this paper we investigate the feasibility of a simple linear transition model.

D. Bayesian Formulation

Integrating the musical score model (prior distribution of the musical notes $\mathbf{Z} = \{\mathbf{P}, \mathbf{O}\}$) with the F0 trajectory model (likelihood function of $\mathbf{Z}$ for the vocal F0s $\mathbf{X}$), we formulate an HHSMM with the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\epsilon}, \boldsymbol{\delta}, \sigma\}$ as follows:

$$p(\mathbf{X}, \mathbf{S}, \mathbf{P}, \mathbf{O}, \mathbf{E}, \mathbf{D} | \boldsymbol{\Theta}) = \underbrace{p(\mathbf{S}|\boldsymbol{\pi}) p(\mathbf{P}|\mathbf{S}, \boldsymbol{\phi}) p(\mathbf{O}|\boldsymbol{\lambda})}_{\text{Musical score model}} \cdot \underbrace{p(\mathbf{E}|\boldsymbol{\epsilon}) p(\mathbf{D}|\boldsymbol{\delta}) p(\mathbf{X}|\mathbf{P}, \mathbf{O}, \mathbf{E}, \mathbf{D}, \sigma)}_{\text{F0 trajectory model}}, \tag{12}$$

where the three terms of the musical score model are given by Eqs. (1) and (2), Eqs. (3) and (4), and Eq. (7), respectively,

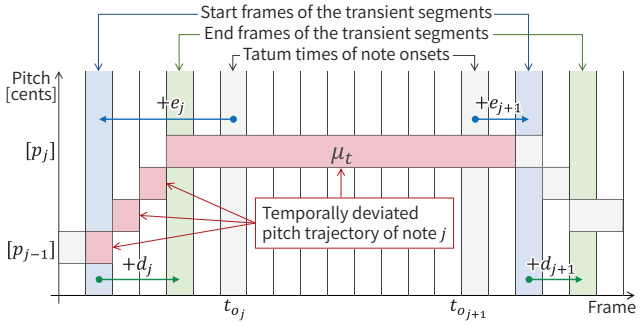


Fig. 6. Temporally deviated pitch trajectory used as Cauchy location parameters. The blue and green vertical boxes represent the start and end frames of the transient segment and the grey vertical boxes represent the tatum times of note onsets. The red boxes represent the temporally deviated pitch trajectory of note j and the grey boxes represent the temporally deviated pitch trajectory of the other notes.

and the three terms of the F0 trajectory model are given by Eq. (8), Eq. (9), and Eq. (10), respectively.

We put conjugate Dirichlet priors as follows:

$$\boldsymbol{\pi}_s \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}_s^{\pi}) \quad (s \in \{0, \ldots, 24\}), \tag{13}$$

$$\bar{\boldsymbol{\phi}}_{rh} \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}_r^{\bar{\phi}}) \quad (r \in \{\text{major, minor}\}, h \in \{0, \ldots, 12\}), \tag{14}$$

$$\boldsymbol{\lambda}_l \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}^{\lambda}) \quad (l \in \{1, \ldots, L\}), \tag{15}$$

$$\boldsymbol{\epsilon} \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}^{\epsilon}), \tag{16}$$

$$\boldsymbol{\delta} \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}^{\delta}), \tag{17}$$

where $\boldsymbol{\gamma}_s^{\pi} \in \mathbb{R}_+^{24}$, $\boldsymbol{\gamma}_r^{\bar{\phi}} \in \mathbb{R}_+^{12}$, $\boldsymbol{\gamma}^{\lambda} \in \mathbb{R}_+^{L}$, $\boldsymbol{\gamma}^{\epsilon} \in \mathbb{R}_+^{e_{\max}-e_{\min}+1}$, and $\boldsymbol{\gamma}^{\delta} \in \mathbb{R}_+^{d_{\max}}$ are hyperparameters. We put a gamma prior on σ as follows:

$$\sigma \sim \mathrm{Gamma}(\gamma_0^{\sigma}, \gamma_1^{\sigma}), \tag{18}$$

where γ_0^{σ} and γ_1^{σ} are the shape and rate parameters of the gamma distribution, which are also hyperparameters.

IV. Posterior Inference

This section explains posterior inference of the latent variables and parameters. The proposed HHSMM can be trained in an unsupervised manner from the vocal F0 trajectory of a target musical piece by sampling the values of the parameters and latent variables to approximate the posterior distribution of those variables (Section IV-A). The HHSMM can also be trained in a semi-supervised manner by using the musical score model that is pretrained by using a large amount of musical scores and that is expected to learn common musical grammar and improve the musical appropriateness of the transcription results. (Section IV-B). The musical notes are finally estimated as the latent variables that maximize the posterior probability of them (Section IV-C). To obtain better results, the parameters are updated simultaneously with the latent variables.

A. Unsupervised Learning

Given an F0 trajectory $\mathbf{X}$ as observed data, our goal is to compute the posterior distribution $p(\mathbf{S}, \mathbf{P}, \mathbf{O}, \mathbf{E}, \mathbf{D}, \boldsymbol{\Theta} | \mathbf{X})$ of the latent variables (the pitches $\mathbf{P}$, onset score times $\mathbf{O}$, onset

deviations $\mathbf{E}$, and transient durations $\mathbf{D}$ of musical notes with the local key $\mathbf{S}$) and the parameters $\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\epsilon}, \boldsymbol{\delta}, \sigma\}$. Since the posterior distribution cannot be computed analytically, we use a Gibbs sampling method with efficient forward-backward procedures and a Metropolis-Hastings (MH) step. The initial values of $\mathbf{Q} = \{\mathbf{P}, \mathbf{O}, \mathbf{E}, \mathbf{D}\}$ are given by quantizing $\mathbf{X}$ on the semitone and tatum grids by the majority vote method. The initial values of $\boldsymbol{\Theta}$ are drawn from Eqs. (13), (14), (15), (16), (17), and (18). Then, the following three steps are iterated until the likelihood is fully maximized.

1) Obtain $\mathbf{S}$ from $p(\mathbf{S}|\mathbf{Q}, \boldsymbol{\Theta}, \mathbf{X})$ with forward filtering-backward sampling.
2) Obtain $\mathbf{Q}$ from $p(\mathbf{Q}|\mathbf{S}, \boldsymbol{\Theta}, \mathbf{X})$ with forward filtering-backward sampling.
3) Obtain $\boldsymbol{\Theta}$ from $p(\boldsymbol{\Theta}|\mathbf{S}, \mathbf{Q}, \mathbf{X})$ with Gibbs sampling and MH sampling.

Detailed algorithms for the steps 1), 2), and 3) are described in Appendix A.

B. Semi-supervised Learning

An effective way of improving the performance of AST is to estimate the parameters of the musical score model from existing musical scores (monophonic note sequences with key annotations) in advance. Let $\hat{\mathbf{S}}$ and $\hat{\mathbf{Z}} = \{\hat{\mathbf{P}}, \hat{\mathbf{O}}\}$ denote local keys, pitches, and onset score times in training data, which are defined in the same way as $\mathbf{S}$ and $\mathbf{Z} = \{\mathbf{P}, \mathbf{O}\}$ of a target piece (Section III-B). Given $\hat{\mathbf{S}}$ and $\hat{\mathbf{Z}}$, the initial and transition probabilities of pitch classes $\bar{\boldsymbol{\phi}}$ are obtained by normalizing the count data $\mathbf{c}^{\bar{\boldsymbol{\phi}}}_{rh}$ obtained from $\hat{\mathbf{S}}$ and $\hat{\mathbf{Z}}$. Similarly, the onset transition probabilities $\boldsymbol{\lambda}$ are obtained from $\hat{\mathbf{O}}$. The initial and transition probabilities of local keys $\boldsymbol{\pi}$ are not trained because the key transitions tend to be unique to each musical piece and are learned in an unsupervised manner. Keeping $\bar{\boldsymbol{\phi}}$ and $\boldsymbol{\lambda}$ fixed, the other parameters and latent variables are estimated for a target piece in a semi-supervised manner.

C. Posterior Maximization

Our final goal is to obtain the optimal values of $\mathbf{S}$, $\mathbf{Q}$, and $\boldsymbol{\Theta}$ that maximize the posterior probability $p(\mathbf{S}, \mathbf{Q}, \boldsymbol{\Theta}|\mathbf{X})$. First, we choose the best samples of $\mathbf{S}$, $\mathbf{Q}$, and $\boldsymbol{\Theta}$ that maximize $p(\mathbf{S}, \mathbf{Q}, \boldsymbol{\Theta}|\mathbf{X})$ in the Gibbs sampling described in IV-A. Then, the following three steps are iterated until convergence.

1) Obtain $\mathbf{S}$ that maximizes $p(\mathbf{S}|\mathbf{Q}, \boldsymbol{\Theta}, \mathbf{X})$ with Viterbi decoding on the upper-level chain of $\mathbf{S}$.
2) Obtain $\mathbf{Q}$ that maximizes $p(\mathbf{Q}|\mathbf{S}, \boldsymbol{\Theta}, \mathbf{X})$ with Viterbi decoding on the lower-level chain of $\mathbf{Q}$.
3) Obtain $\boldsymbol{\Theta}$ that maximizes $p(\boldsymbol{\Theta}|\mathbf{S}, \mathbf{Q}, \mathbf{X})$.

We empirically confirmed that a few iterations are sufficient to reach convergence. In the Viterbi algorithm in step 2 above, weighting factors β^{ϕ}, β^{λ}, β^{ϵ}, β^{δ}, and β^{χ} are introduced in the forward calculations to balance the individual sub-models. A penalization term $\exp[\beta^{o}/(o_{j+1} - o_j)]$ for long durations $o_{j+1} - o_j$ with a weighting factor β^{o} is also introduced in the forward calculations to suppress the frequent occurrence of long notes. Detailed algorithms for 1), 2), and 3) are described in Appendix B.

V. Evaluation

We conducted comparative experiments to evaluate the performance of the proposed method for AST. We investigated the effectiveness of the pretraining method in comparison with the unsupervised and semi-supervised learning methods (Section V-B1) based on the learning configurations described in Section V-A3. We also examined the contribution of the individual sub-models by ablating each of them (Sections V-B2, V-B3, and V-B4) based on the model configurations described in Section V-A4. To confirm the improvement of the performance of the proposed method, we conducted a comparative experiment using conventional methods (Section V-C1). To investigate the performance of the overall system that takes a music signal as input and outputs a musical score, we also tested the proposed method for F0 trajectories and tatum times automatically estimated from music signals (Section V-C2).

A. Experimental Conditions

1) Datasets: From the RWC Music Database [32], we used 63 popular songs in 4/4 time and that satisfy the requirements mentioned in Section III-A. We verified the correctness of ground-truth annotations [33] of musical notes and beat times. For input vocal F0 trajectories, we used the ground-truth data in most experiments and estimated data obtained by the method in [13] for some experiments. In both cases, the ground-truth unvoiced regions were used to eliminate the influence of the performance of vocal activity detection (VAD). Similarly, for the 16th-note-level tatum times, we used the ground-truth data in most experiments and estimated data obtained by a neural beat tracking method [8] in some experiments. To prepare ground-truth scores used for evaluating the accuracy of transcribed notes, we used MIDI files in which the onset and offset times of each note are adjusted to corresponding ground-truth beat times.

2) Hyperparameters: The Dirichlet priors on the initial key probabilities $\boldsymbol{\pi}_0$, the onset transition probabilities $\boldsymbol{\lambda}$, the onset time deviation probabilities $\boldsymbol{\epsilon}$, and the transient duration probabilities $\boldsymbol{\delta}$ given by Eqs. (13), (15), (16), and (17) were set to uniform distributions, *i.e.*, $\boldsymbol{\gamma}^{\pi}_0$, $\boldsymbol{\gamma}^{\lambda}$, $\boldsymbol{\gamma}^{\epsilon}$, and $\boldsymbol{\gamma}^{\delta}$ were set to all-one vectors. The Dirichlet priors on the key transition probabilities $\boldsymbol{\pi}_s$ ($s \in \{1, \ldots, 24\}$) given by Eq. (13) were set as $\boldsymbol{\gamma}^{\pi}_s = [1, 1, \cdots, 100, \cdots, 1]^{\mathrm{T}}$ (only the s-th element takes 100) to favor self-transition. The Dirichlet priors on the initial probabilities of pitch classes $\bar{\boldsymbol{\phi}}_{r0}$ given by Eq. (14) and the transition probabilities of pitch classes $\bar{\boldsymbol{\phi}}_{rh}$ ($r \in \{\text{major, minor}\}, h \in \{1, \ldots, 12\}$) were set as $\boldsymbol{\gamma}^{\bar{\boldsymbol{\phi}}}_{\text{major}} = [10, 1, 10, 1, 10, 10, 1, 10, 1, 10, 1, 10]^{\mathrm{T}}$ and $\boldsymbol{\gamma}^{\bar{\boldsymbol{\phi}}}_{\text{minor}} = [10, 1, 10, 10, 1, 10, 1, 10, 10, 1, 10, 1]^{\mathrm{T}}$ to favor the seven pitch classes on the C major and minor scales, respectively. The gamma prior on σ in Eq. (18) was set as $a^{\sigma}_0 = a^{\sigma}_1 = 1$. Assuming that keys tend to change infrequently, the s-th element of $\boldsymbol{\gamma}^{\pi}_s$ was set to a large value (100). Because non-diatonic notes are often used, $\boldsymbol{\gamma}^{\bar{\boldsymbol{\phi}}}_{\text{major}}$ and $\boldsymbol{\gamma}^{\bar{\boldsymbol{\phi}}}_{\text{minor}}$ were set to smaller values (10). Optimization of these hyperparameters is left as future work.

For a model M3 in Table I the weighting factors β^{ϕ}, β^{λ}, and β^{χ} were determined by Bayesian optimization [11] as

$\beta^{\phi} = 18.9$, $\beta^{\lambda} = 49.6$, and $\beta^{\chi} = 5.1$. The weighting factors β^{ϵ} and β^{δ} were determined by grid search and set as $\beta^{\epsilon} = 20.0$ and $\beta^{\delta} = 10.0$. The weighting factor β^{o} of the duration penalty term was set to $\beta^{o} = 50$, which was experimentally selected from $\{1, 5, 10, 50, 100, 500, 1000\}$ so that the performances of M3 and M4 were maximized. Since the forward-backward algorithms (Appendices A-2 and B-2) are defined in a huge product space $\bar{q}_n = \{\bar{p}_n, \bar{o}_n, \bar{e}_n, \bar{d}_n\}$, the range of pitches considered was limited as follows:

$$\bar{p}_n \in \bigcup_{i=n-1}^{n+1} \left\{ p_i^{\mathrm{Maj}} - 1, p_i^{\mathrm{Maj}}, p_i^{\mathrm{Maj}} + 1 \right\}, \tag{19}$$

where p_n^{Maj} is the pitch estimated by the majority vote method between tatums $n-1$ and n. The pitch-range constraint might prevent the proposed method from estimating some correct notes. However, it is difficult to recover the correct notes from an erroneous F0 trajectory that is far from the ground-truth pitch sequence. The pitch-range constraint is thus effective for reducing the computational complexity of the proposed method without much damaging the performance of note estimation.

3) Learning Configurations: The unsupervised and semi-supervised schemes (Section IV-A and Section IV-B) were used for estimating the initial and transition probabilities of pitch classes $\bar{\phi}$ and the onset transition probabilities $\boldsymbol{\lambda}$. In the unsupervised scheme $\bar{\phi}$ and/or $\boldsymbol{\lambda}$ were learned from only the vocal F0 trajectory of a target song. In the semi-supervised scheme $\bar{\phi}$ and/or $\boldsymbol{\lambda}$ were estimated as follows:

- [L1] Pitch transition learning: $\bar{\phi}$ were learned from 90 popular songs with no overlapped sung notes except for a target song in the RWC Music Database [32].
- [L2] Onset transition learning: $\boldsymbol{\lambda}$ were estimated in advance from a corpus of rock music [34].

4) Model Configurations: The four main components of the proposed method, *i.e.*, the local key model (Section III-B1), rhythm model (Section III-B3), temporal deviation model (Section III-C1), and note duration penalty (Section IV-C) were evaluated separately. More specifically, we examined the performance of AST when each component was included or not included as follows:

- [T1] Key modeling: When this function was enabled, the key transition probabilities $\boldsymbol{\pi}$ were estimated. Otherwise, the number of local keys was set to 1, *i.e.*, $\pi_{0,1} = \pi_{1,1} = 1$.
- [T2] Rhythm modeling: When this function was enabled, the onset transition probabilities $\boldsymbol{\lambda}$ were estimated. Otherwise, $\boldsymbol{\lambda}$ were set to the same value, *i.e.*, $\lambda_{l,l'} = 1$.
- [T3] Temporal deviation modeling: When this function was enabled, the onset deviations $\mathbf{E}$ and the transient durations $\mathbf{D}$ were estimated and we set $[e_{\min}, e_{\max}] = [-5, 5]$ and $d_{\max} = 10$. Otherwise, $\mathbf{E}$ and $\mathbf{D}$ were not considered, *i.e.*, $e_j = 0$ and $d_j = 1$.
- [T4] Note duration penalization: When this function was enabled, the penalty term $f(o_n)$ was introduced in the Viterbi decoding. Otherwise, $f(o_n)$ was not used.

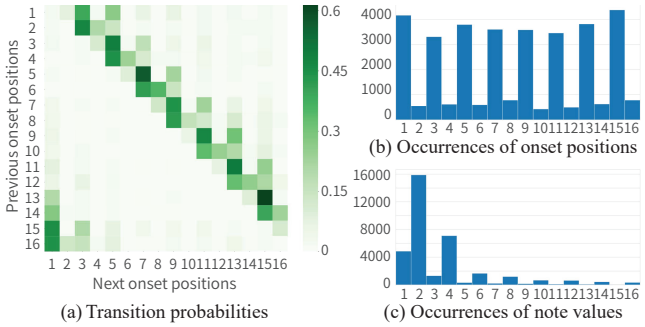


Fig. 7. Pretrained transition probabilities between the 16th-note-level tatum positions.

5) Evaluation Measures: The performance of AST was evaluated in terms of the frame-level and note-level measures. The frame-level accuracy, $\mathcal{A}$, is defined as follows:

$$\mathcal{A} = \frac{T_{\mathrm{correct}}}{T_{\mathrm{gt}}}, \tag{20}$$

where T_{gt} is the total number of voiced frames in the beat-adjusted MIDI files and T_{correct} is the total number of voiced frames whose semitone-level pitches were correctly estimated. The note-level measures including the precision rate $\mathcal{P}$, the recall rate $\mathcal{R}$, and the F-measure $\mathcal{F}$ are defined as follows:

$$\mathcal{P} = \frac{J_{\mathrm{correct}}}{J_{\mathrm{output}}}, \quad \mathcal{R} = \frac{J_{\mathrm{correct}}}{J_{\mathrm{gt}}}, \quad \mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}}, \tag{21}$$

where J_{gt} is the total number of musical notes in the ground-truth scores, J_{output} is the total number of musical notes in the transcribed scores, and J_{correct} is the total number of correctly-estimated musical notes. We used three criteria for judging that a musical note is correctly estimated [35]:

- [C1] Pitch, onset, and offset match: The semitone-level pitch and the onset and offset score times were all estimated correctly.
- [C2] Pitch and onset match: The semitone-level pitch and the onset score time were estimated correctly.
- [C3] Onset match: The onset score time was estimated correctly.

When the tatum times were estimated with the beat tracking method [8], we used error tolerance for onset and offset times to absorb slight differences between the estimated and ground-truth tatum times. More specifically, an estimated onset time was judged as correct if it was within 50 ms from the ground-truth onset time. An estimated offset time was judged as correct if it was within 50 ms or 20% of the duration of the ground-truth note from the ground-truth offset time.

B. Experimental Results

The experimental results obtained with the different learning and model configurations are shown in Table I. In this experiment, the ground-truth vocal F0 trajectories and tatum times were given as input to the proposed method.

TABLE I
PERFORMANCE OF THE PROPOSED METHOD WITH DIFFERENT LEARNING AND MODEL CONFIGURATIONS.

	L1 Pitch trans. learning	L2 Onset trans. learning	T1 Key modeling	T2 Rhythm modeling	T3 Temp. dev. modeling	T4 Note dur. penalization	$\mathcal{A}$	C1 Pitch, onset, offset $\mathcal{P}$	C1 $\mathcal{R}$	C1 $\mathcal{F}$	C2 Pitch, onset $\mathcal{P}$	C2 $\mathcal{R}$	C2 $\mathcal{F}$	C3 Onset $\mathcal{P}$	C3 $\mathcal{R}$	C3 $\mathcal{F}$
M1			✓	✓			71.4	21.4	16.6	18.6	40.7	30.5	34.6	52.8	39.1	44.6
M2	✓		✓	✓			71.5	21.3	16.3	18.3	41.0	30.3	34.6	53.7	39.3	45.0
M3		✓	✓	✓			73.5	**28.7**	24.0	26.0	**47.1**	38.9	42.3	**62.6**	51.5	56.2
M4	✓	✓	✓	✓			73.6	28.2	23.3	25.3	**47.1**	38.4	42.0	**62.6**	51.0	55.8
M5							71.7	21.7	19.2	20.2	39.9	35.1	37.1	55.2	48.8	51.5
M6			✓				72.9	22.2	19.8	20.8	40.8	36.1	38.0	55.0	49.0	51.5
M7		✓		✓			72.6	27.7	22.9	24.9	46.2	37.6	41.2	62.5	50.8	55.6
M8		✓	✓	✓	✓		73.5	27.3	23.8	25.3	45.5	39.2	41.9	61.2	52.8	56.3
M9		✓	✓	✓		✓	**73.7**	28.4	**26.5**	**27.3**	44.6	41.3	**42.6**	61.3	57.0	58.7
M10		✓	✓	✓	✓	✓	73.1	27.5	26.4	26.8	43.8	**41.9**	**42.6**	60.5	**58.1**	**58.9**
M11	✓	✓	✓	✓	✓	✓	73.1	27.1	25.1	25.9	44.1	40.6	42.0	60.8	56.3	58.1

TABLE II
PERFORMANCE OF THE CONVENTIONAL AND PROPOSED METHODS.

	$\mathcal{A}$	C1 $\mathcal{P}$	C1 $\mathcal{R}$	C1 $\mathcal{F}$	C2 $\mathcal{P}$	C2 $\mathcal{R}$	C2 $\mathcal{F}$	C3 $\mathcal{P}$	C3 $\mathcal{R}$	C3 $\mathcal{F}$
Majority-vote method	54.0	10.1	19.1	13.1	21.2	42.6	28.0	37.3	77.7	49.9
BS-HMM [36]	64.2	9.3	9.4	9.2	23.9	23.7	23.5	37.6	37.7	37.2
M3 (proposed method)	**73.5**	**34.7**	**32.4**	**33.3**	**49.6**	**46.1**	**47.4**	**64.5**	59.9	61.6

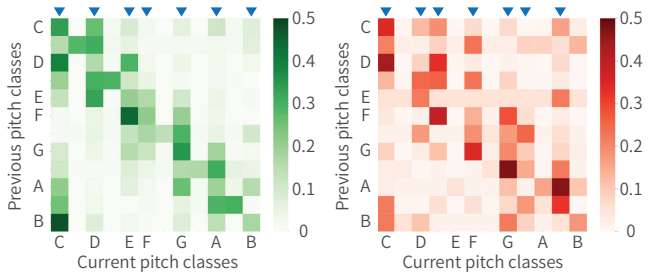
TABLE III
PERFORMANCE OF THE PROPOSED METHOD BASED ON E0 ESTIMATION AND/OR TATUM DETECTION.

	F0 trajectory	Tatum times	$\mathcal{A}$	C1 $\mathcal{P}$	C1 $\mathcal{R}$	C1 $\mathcal{F}$	C2 $\mathcal{P}$	C2 $\mathcal{R}$	C2 $\mathcal{F}$	C3 $\mathcal{P}$	C3 $\mathcal{R}$	C3 $\mathcal{F}$
M3	Ground-truth	Ground-truth	**73.5**	**28.7**	24.0	**26.0**	**47.1**	38.9	**42.3**	**62.6**	51.5	**56.2**
M3-1	[13]	Ground-truth	73.0	28.1	**24.4**	25.9	46.4	**39.7**	42.5	61.3	**52.5**	56.1
M3-2	Ground-truth	[8]	71.2	25.7	21.5	23.3	43.5	35.9	39.0	58.3	47.8	52.1
M3-3	[13]	[8]	70.7	25.9	22.7	24.1	43.6	37.7	40.1	57.8	49.8	53.2

1) Learning Configurations: We evaluated the effectiveness of the semi-supervised learning strategy by comparing the performances for the four different learning configurations M1, M2, M3, and M4 in Table I. The accuracies for M3 (73.5%) and M4 (73.7%) were better than those for M1 (71.4%) and M2 (71.5%). This indicates that the pretraining of the onset transition probabilities $\boldsymbol{\lambda}$ was effective for improving the performance of AST. Examples of $\boldsymbol{\lambda}$ are shown in Fig. 7. The probabilities of transitions to the 8th-note-level tatum grids were larger than those to the other positions (Figs. 7-(a) and 7-(b)) and the 8th notes frequently appeared in the date used for the pretraining (Fig. 7-(c)). This enables the proposed method to output scores with natural rhythms.

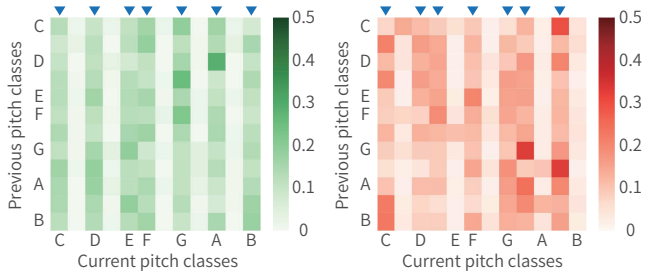
The accuracy for M1 (71.4%) was almost the same as that for M2 (71.5%) and that for M3 (73.5%) was almost the same as that for M4 (73.6%). This indicates that the pretraining of the transition probabilities of pitch classes $\bar{\phi}$ did not much improve the performance of AST. The precision rates, recall rates, and F-measures for M2 and M4 were slightly worse than those for M1 and M3 especially in terms of C1, respectively. As shown in Fig. 8, when the transition probabilities of pitch classes were pretrained, pitches were likely to continue or change to near pitches. As shown in Figs. 9 and 10, in contrast, there was no clear bias in the transition probabilities estimated by the unsupervised learning method or the posterior maximization method because of the effect of the prior distribution. Estimation errors caused by using the pretrained transition probabilities are shown in Fig. 11, where the ground-truth key was D minor in both examples. In the example 1, M3 failed to estimate the correct key, but succeeded in estimating the correct notes on the scale of the estimated key. In contrast, M4 failed to estimate the correct key and two estimated notes were out of the scale of the estimated key. M4 also failed to estimate even the note on the scale of the estimated key because vocal F0s corresponding to the note were represented as frequency deviations. In the example 2, M3 succeeded in estimating the correct note on the scale of the incorrectly estimated key. M4 estimated a key as F major (a relative major key of D minor), but incorrectly estimated the note because vocal F0s corresponding to the note were represented as frequency deviations. As in these examples, incorrect key estimation and/or incorrectly represented F0s sometimes led to incorrect note estimation.

2) Key and Rhythm Modeling: We evaluated the effectiveness of the key and rhythm modeling by comparing the performances for M3, M5, M6, and M7 in Table I. Note that the pretrained onset transition probabilities $\boldsymbol{\lambda}$ were not used for M5 and M6 that do not model rhythms. The accuracy for M3 (73.6%) was better than those for M5 (71.7%), M6 (72.9%), and M7 (72.6%). This indicates that the musical score model including the key and rhythm models was effective for improving the performance of AST. While the accuracy for M6



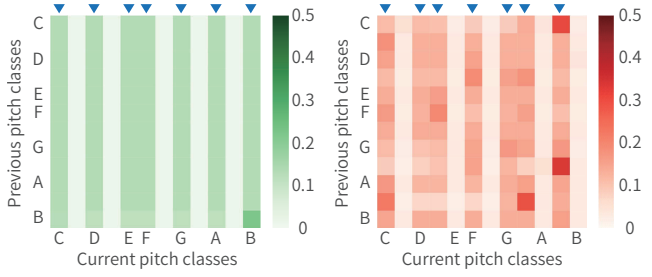
(a) Transition probabilities between pitch classes in C major (b) Transition probabilities between pitch classes in C minor

Fig. 8. Pretrained transition probabilities between the 12 pitch classes under the major and minor diatonic scales.



(a) Transition probabilities between pitch classes in C major (b) Transition probabilities between pitch classes in C minor

Fig. 9. Transition probabilities between the 12 pitch classes estimated by the unsupervised learning method (Section IV-A).



(a) Transition probabilities between pitch classes in C major (b) Transition probabilities between pitch classes in C minor

Fig. 10. Transition probabilities between the 12 pitch classes estimated by the posterior maximization method (Section IV-C).

(72.9%) was almost the same as that for M7 (72.6%), the precision rates, recall rates, and F-measures for M7 were much better than those for M6. This indicates that the rhythm model was more effective than the key model for AST.

3) Temporal Deviation Modeling: We evaluated the effectiveness of the temporal deviation modeling by comparing the performances for M3 and M8 in Table I. In most criteria, the performances for M8 were worse than those for M3. This was mainly due to the large deviations of a vocal F0 trajectory that were hard to represent precisely. The positive effect of the temporal deviation modeling was shown in Fig. 12. In these pieces, M8 successfully represented the transient segments of vocal F0s. In the left piece, M8 correctly estimated a musical note regarded as frequency deviations by M3. The temporal deviation model of M8, however, often degraded the performance, as shown in Fig. 13. In the top pieces, the temporal deviation model overfit *overshoot* and *preparation* in vocal F0 trajectories. Although we had expected the overshoot and preparation to be captured as frequency deviations following the Cauchy distribution, the overfitting, in fact, caused inaccurate pitch estimation. In the bottom-left piece, an extra note was inserted because the transient segment longer than the tatum interval was represented by the temporal deviations of the two successive tatums. In the lower-right piece, the proposed method failed to detect the correct onset because the transition segment was represented as the temporal deviation of the previous tatum before the correct tatum. Treatment of such slower F0 transitions should be included in future work.

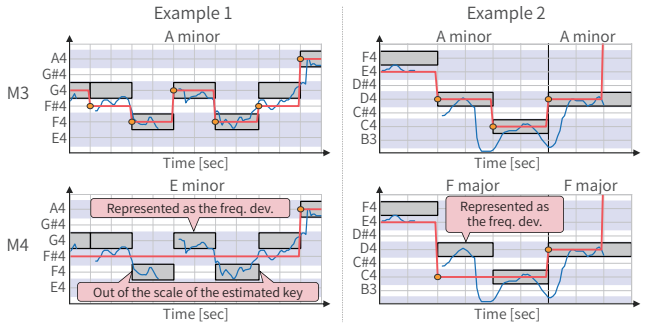


Fig. 11. Estimation errors caused by using the pretrained initial and transition probabilities of pitch classes. The pale blue backgrounds indicate the diatonic scales of estimated keys and the gray boxes indicate ground-truth musical notes. The blue and red lines indicate vocal F0s and estimated musical notes, respectively. The orange dots indicate estimated note onsets. The gray grids indicate tatum times and semitone-level pitches. The red balloons indicate the ground-truth notes that the proposed method failed to estimate. The estimated keys are illustrated in the figure, and the ground-truth key in both examples is D minor.

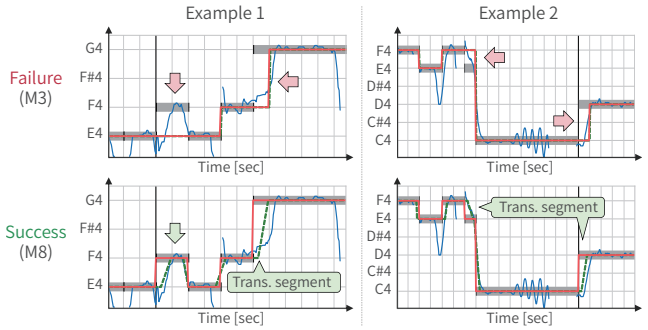


Fig. 12. Positive effects of temporal deviation modeling (cf. Fig. 11). The green lines estimated F0s with temporal deviations. The red arrows indicate estimation errors and the green arrows and balloons indicate correct notes obtained by modeling temporal deviations.

As shown in Fig. 14, the categorical distribution of the onset time deviations $\mathbf{E}$ and that of the transient durations $\mathbf{D}$ estimated by the unsupervised learning method (Section IV-A) were found to have complicated shapes. In future work, flexible parametric models (*e.g.*, Poisson mixture models) that have fewer parameters than the categorical distribution could be used for improving the efficiency of training the distributions of the temporal deviations $\mathbf{E}$ and $\mathbf{D}$.

4) Note Duration Penalization: We evaluated the effectiveness of introducing the penalty term $f(o_n)$ by comparing the

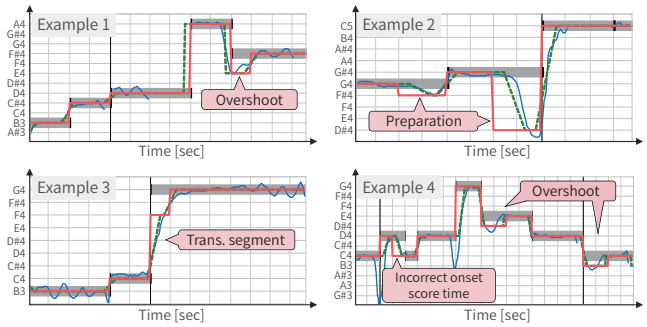


Fig. 13. Negative effects of temporal deviation modeling (cf. Fig. 12). The red balloons indicate estimation errors.

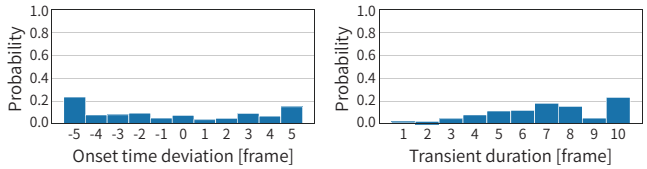


Fig. 14. The categorical distribution of the onset time deviations $\mathbf{E}$ and that of the transient durations $\mathbf{D}$ estimated in the unsupervised learning method (Section IV-A).

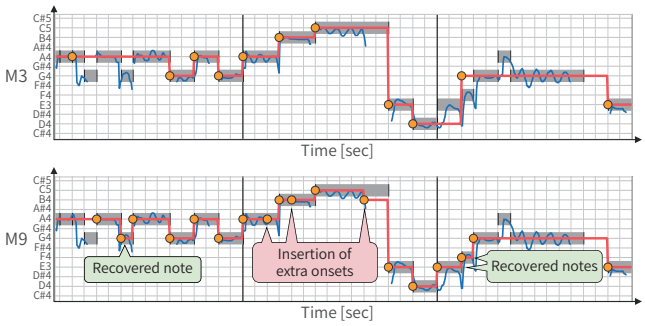


Fig. 15. Positive and negative effects of duration penalization (cf. Fig. 11). The green and red balloons indicate improved and deteriorated parts.

performances for the two different models M3 and M9 in Table I. The accuracy for M9 (73.7%) was slightly better than that for M9 (73.5%). In terms of C1, the precision rate for M9 (28.4%) was slightly worse than that for M3 (28.7%), whereas the recall rate for M9 (26.5%) was better than that for M3 (24.0%). As shown in Fig. 15, M9 successfully estimated shorter notes that were not detected by M3. On the other hand, M9 mistakenly split longer notes correctly estimated by M3 into several shorter notes. A possible way of mitigating this trade-off problem is to extend the proposed model to deal with both pitch and volume trajectories of singing voice. This is expected to improve the performance of onset detection.

We evaluated the effectiveness of jointly using the temporal deviation modeling and the note duration penalization by comparing the performances for M9 an M10 in Table I. As described in Section V-B3, the temporal deviation modeling decreased the accuracy from 73.7% (M9) to 73.1% (M10). We also checked the effectiveness of introducing both models into M4. Although the accuracy for M11 and that for M10 were same, the precision and recall rates and F-measures for M11 were slightly worse than those for M10. This indicates that the pretrained probabilities $\bar{\phi}$ have little effect when using the temporal deviation modeling and note duration penalization.

C. Further Investigations

1) Comparison with Conventional Methods: We compared the performance obtained by M3 with those obtained by the straightforward majority-vote method and a statistical method based on a beat-synchronous HMM (BS-HMM) [36]. Since these conventional methods can only estimate a semitone-level pitch at each tatum interval, successive tatum intervals were concatenated to form a single note if they had the same pitch. Similarly, successive notes included in the ground-truth data and those estimated by the proposed method were concatenated to form a single note if they had the same pitch. To deal with unvoiced regions in vocal F0 trajectories, the BS-HMM was extended in the same way as the proposed method. Note that the majority-vote method can handle the unvoiced regions and estimate rest notes for those regions.

Experimental results are shown in Table II. The accuracy obtained by the proposed method (73.5%) was better than those obtained by the majority-vote method (54.0%) and the BS-HMM (64.2%). Considering that the use of the musical score model is the main difference between the proposed method and the others, we can say that the musical score model is effective for improving the performance of AST. In terms of C3, only the recall rate obtained by the majority-vote method (77.7%) was better than those obtained by the other methods because many extra short notes were obtained.

2) Impacts of F0 Estimation and Tatum Detection: To evaluate the impacts of vocal F0 estimation and tatum time detection on the performance of AST, we tested M3 by using the ground-truth and estimated data as follows:

- *F0 trajectory*: We used the ground truth data in [33] and estimated data obtained by [13].
- *Tatum times*: We used the ground truth data in [33] and estimated data detected by [8].

As shown in Table III, the accuracy obtained by M3 (73.5%) using the ground-truth data was better than those obtained by M3-1 (73.0%), M3-2 (71.2%), and M3-3 (70.7%) using the estimated data. The impact of tatum detection was larger than that of F0 estimation and the F0 estimation slightly degraded the performance of the proposed method. To avoid such negative effects, we plan to develop a method that directly and jointly estimates a sequence of musical notes with tatum times from music audio signals.

VI. Conclusion

This paper presented a statistical AST method that estimates a sequence of musical notes from a vocal F0 trajectory. Our method is based on an HHSMM that combines a musical score model with a vocal F0 model to represent the hierarchical generative process of a vocal F0 trajectory from a note sequence determined by musical knowledge. We derived an efficient Bayesian inference algorithm based on the Gibbs and MH samplings for solving the inverse problem. The experimental results showed that the key and rhythm models included

in the musical score model were effective for improving the performance of AST. The temporal deviation model was not always effective because it was often difficult to discriminate temporal deviations and frequency deviations.

One of future research directions is to extend the proposed method for directly and jointly estimating musical notes and tatum times from music audio signals. This could be realized by integrating the proposed method with an acoustic model that generates a music spectrogram from a vocal F0 trajectory in a hierarchical Bayesian manner.

APPENDIX A
DETAILED ALGORITHM FOR UNSUPERVISED LEARNING

1) Sampling Local Keys: In the forward step, a forward message $\alpha(s_m)$ is calculated recursively as follows:

$$\alpha(s_0) = p(p_0, s_0) = p(p_0|s_0)p(s_0) = \phi_{s_0,0,p_0}\pi_{0,s_0}, \tag{22}$$

$$\begin{aligned}\alpha(s_m) &= p(p_{0:j_{m+1}-1}, s_m) \\ &= \sum_{s_{m-1}} p(s_m|s_{m-1})\alpha(s_{m-1}) \prod_{j=j_m}^{j_{m+1}-1} p(p_j|p_{j-1}, s_m) \\ &= \sum_{s_{m-1}} \pi_{s_{m-1}s_m}\alpha(s_{m-1}) \prod_{j=j_m}^{j_{m+1}-1} \phi_{s_m p_{j-1} p_j},\end{aligned} \tag{23}$$

where j_m denotes the index of the first note in measure m.

In the backward step, the local keys $\mathbf{S}$ are sampled from a conditional distribution given by

$$p(\mathbf{S}|\mathbf{P}) = p(s_M|\mathbf{P}) \prod_{m=0}^{M-1} p(s_m|s_{m+1:M}, \mathbf{P}). \tag{24}$$

More specifically, local keys $s_{0:M}$ are sampled in the backward order as follows:

$$s_M \sim p(s_M|\mathbf{P}) \propto \alpha(s_M), \tag{25}$$

$$s_m \sim p(s_m|s_{m+1:M}, \mathbf{P}) \propto \pi_{s_m s_{m+1}}\alpha(s_m). \tag{26}$$

2) Sampling Note-Level Variables: Given the local keys $\mathbf{S}$ and the F0 trajectory $\mathbf{X}$, we aim to jointly update $\mathbf{Q} = \{\mathbf{P}, \mathbf{O}, \mathbf{E}, \mathbf{D}\}$ by using a forward filtering-backward sampling algorithm on the tatum grids. We define a forward message $\bar{\alpha}(\bar{q}_n)$ w.r.t. a tuple $\bar{q}_n = \{\bar{p}_n, \bar{o}_n, \bar{e}_n, \bar{d}_n\}$ (Fig. 16), where $\bar{p}_n$ and $\bar{o}_n$ indicate the pitch and onset score time of the note whose offset score time is given by n, and $\bar{e}_n$ and $\bar{d}_n$ respectively indicate the onset time deviation and transient duration of the note whose onset score time is given by n. The onset and offset times of the musical note whose offset time is n are thus given by $t_{\bar{o}_n} + \bar{e}_{\bar{o}_n}$ and $t_n + \bar{e}_n - 1$. We formally write the emission probability of F0s in this time span as follows:

$$\chi(\bar{q}_n) = \prod_{t=t_{\bar{o}_n}+\bar{e}_{\bar{o}_n}}^{t_n+\bar{e}_n-1} \mathrm{Cauchy}(x_t|\mu_t, \sigma), \tag{27}$$

where μ_t is given by the piecewise linear trajectory given by Eq. (11) as follows:

$$\mu_t = \begin{cases} [\bar{p}_{\bar{o}_n}] + ([\bar{p}_n] - [\bar{p}_{\bar{o}_n}])(t - (t_{\bar{o}_n} + \bar{e}_{\bar{o}_n}))/\bar{d}_{\bar{o}_n} \\ \qquad (t \in [t_{\bar{o}_n} + \bar{e}_{\bar{o}_n}, t_{\bar{o}_n} + \bar{e}_{\bar{o}_n} + \bar{d}_{\bar{o}_n})), \\ [\bar{p}_n] \quad (t \in [t_{\bar{o}_n} + \bar{e}_{\bar{o}_n} + \bar{d}_{\bar{o}_n}, t_n + \bar{e}_n)). \end{cases} \tag{28}$$

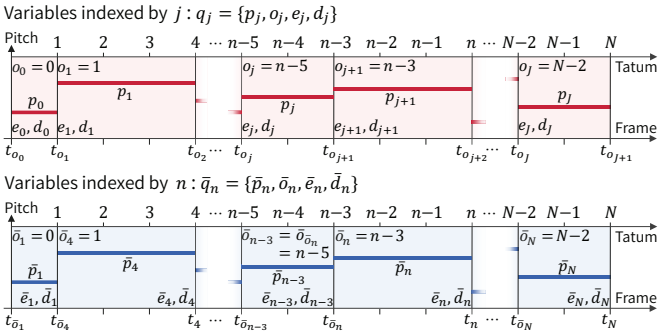


Fig. 16. A relationship between the variables $q_j = \{p_j, o_j, e_j, d_j\}$ and $\bar{q}_n = \{\bar{p}_n, \bar{o}_n, \bar{e}_n, \bar{d}_n\}$.

The variable $\bar{q}_n$ is indexed by tatum n (unlike note j) to enable estimation by a hidden semi-Markov model whereby the number of notes and the onset score time of each note are obtained as a result [37].

In the forward step, a forward message $\bar{\alpha}(\bar{q}_n)$ is calculated recursively as follows:

$$\bar{\alpha}(\bar{q}_1) = p(\bar{q}_1|\mathbf{S}, \mathbf{X}) = \phi_{s_0,0,\bar{p}_1}, \tag{29}$$

$$\begin{aligned}\bar{\alpha}(\bar{q}_n) &= p(x_{0:t_n+\bar{e}_n-1}, \bar{q}_n|\mathbf{S}, \mathbf{X}) \\ &= \sum_{\bar{q}_{\bar{o}_n}} p(x_{0:t_n+\bar{e}_n-1}, \bar{q}_{\bar{o}_n}, \bar{q}_n|\mathbf{S}, \mathbf{X}) \\ &= p(\bar{e}_n)p(\bar{d}_n) \sum_{\bar{q}_{\bar{o}_n}} p(n|\bar{o}_n)p(\bar{p}_n|\bar{p}_{\bar{o}_n}, s_{\bar{o}_n})\chi(\bar{q}_n)\bar{\alpha}(\bar{q}_{\bar{o}_n}) \\ &= \epsilon_{\bar{e}_n}\delta_{\bar{d}_n} \sum_{\bar{q}_{\bar{o}_n}} \lambda_{l_{\bar{o}_n}, l_n}\phi_{s_{\bar{o}_n}, \bar{p}_{\bar{o}_n}, \bar{p}_n}\chi(\bar{q}_n)\bar{\alpha}(\bar{q}_{\bar{o}_n}),\end{aligned} \tag{30}$$

where $s_{\bar{o}_n}$ indicates the local key of a measure including the tatum $\bar{o}_n$.

In the backward step, the variables $\mathbf{Q}$ are sampled from a conditional distribution given by

$$p(\mathbf{Q}|\mathbf{S}, \mathbf{X}) = p(q_J|\mathbf{S}, \mathbf{X}) \prod_{j=0}^{J-1} p(q_j|q_{j+1:J}, \mathbf{S}, \mathbf{X}), \tag{31}$$

where $q_j = \{p_j, o_j, e_j, d_j\}$ is a tuple of the semitone-level pitch, onset score time, onset time deviation, and transient duration of j-th note, respectively (Fig. 16). The variables $\mathbf{Q}$, however, cannot be sampled directly from Eq. (31) because the number of notes J is unknown before sampling notes. Instead of sampling q_j, the variable $\bar{q}_n$ is recursively sampled in the reverse order as follows:

$$\bar{q}_N \sim p(\bar{q}_N|\mathbf{S}, \mathbf{X}) \propto \bar{\alpha}(\bar{q}_N), \tag{32}$$

$$\begin{aligned}\bar{q}_{\bar{o}_n} &\sim p(\bar{q}_{\bar{o}_n}|\bar{q}_n, \mathbf{S}, \mathbf{X}) \\ &\propto \epsilon_{\bar{e}_n}\delta_{\bar{d}_n}\lambda_{\bar{o}_n,\bar{o}_n}\phi_{s_{\bar{o}_n}, \bar{p}_{\bar{o}_n}, \bar{p}_n}\chi(\bar{q}_n)\bar{\alpha}(\bar{q}_{\bar{o}_n}).\end{aligned} \tag{33}$$

As a result of the sampling, J is determined as the number of the sampled tuples.

3) Sampling Model Parameters: Given the latent variables $\mathbf{S}$ and $\mathbf{Q}$, the model parameters $\boldsymbol{\Theta}$ except for σ are sampled

from the conditional posterior distributions as follows:

$$\boldsymbol{\pi}_s \mid \mathbf{S} \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}_i^{\pi} + \mathbf{c}_i^{\pi}), \tag{34}$$

$$\bar{\boldsymbol{\phi}}_{rh} \mid \mathbf{S}, \mathbf{P}, \mathbf{O} \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}_r^{\bar{\phi}} + \mathbf{c}_{rh}^{\bar{\phi}}), \tag{35}$$

$$\boldsymbol{\lambda}_l \mid \mathbf{O} \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}^{\lambda} + \mathbf{c}_l^{\lambda}), \tag{36}$$

$$\boldsymbol{\epsilon} \mid \mathbf{E} \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}^{\epsilon} + \mathbf{c}^{\epsilon}), \tag{37}$$

$$\boldsymbol{\delta} \mid \mathbf{D} \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}^{\delta} + \mathbf{c}^{\delta}), \tag{38}$$

where $\mathbf{c}_s^{\pi} \in \mathbb{R}_+^{24}$, $\mathbf{c}_{rh}^{\bar{\phi}} \in \mathbb{R}_+^{12}$, $\mathbf{c}_l^{\lambda} \in \mathbb{R}_+^{L}$, $\mathbf{c}^{\epsilon} \in \mathbb{R}^{e_{\max}-e_{\min}+1}$, and $\mathbf{c}^{\delta} \in \mathbb{R}_+^{d_{\max}}$ are count data obtained from $\mathbf{S}$ and $\mathbf{Q}$. More specifically, c_{0s}^{π} indicates the number of times that $s_0 = s$ is satisfied, $c_{ss'}^{\pi}$ indicates the number of transitions from key s to key s', $c_{r0h}^{\bar{\phi}}$ indicates the number of times that $\mathrm{type}(s_0) = r$ and $\deg(s_0, p_0) = h$ are both satisfied, $c_{rhh'}^{\bar{\phi}}$ indicates the number of transitions from a pitch degree h to a pitch degree h' under a key type r, $c_{ll'}^{\lambda}$ indicates the number of transitions from a tatum position l to a tatum position l', c_e^{ϵ} indicates the number of onset time deviations taking e, and c_d^{δ} indicates the number of transient durations taking d.

To update σ, we use a MH algorithm with a random-walk proposal distribution as follows:

$$q(\sigma^* | \sigma) = \mathrm{Gamma}(\sigma^* | \sigma, 1), \tag{39}$$

where σ is a current sample and σ^* is a proposal. The proposal σ^* is accepted as the next sample with the probability given by

$$\mathcal{A}(\sigma^*, \sigma) = \min\left(\frac{\mathcal{L}(\sigma^*) q(\sigma | \sigma^*)}{\mathcal{L}(\sigma) q(\sigma^* | \sigma)}, 1\right), \tag{40}$$

where $\mathcal{L}(\sigma)$ is the likelihood function of σ given by

$$\mathcal{L}(\sigma) = \mathrm{Gamma}(\sigma | \gamma_0^{\sigma}, \gamma_1^{\sigma}) \prod_{j=1}^{J} \prod_{t=t_{o_j}+e_{o_j}}^{t_{o_{j+1}}+e_{o_{j+1}}-1} \mathrm{Cauchy}(x_t | \mu_t, \sigma). \tag{41}$$

APPENDIX B
DETAILED ALGORITHM FOR POSTERIOR MAXIMIZATION

1) Estimating Local Keys: In the forward step, a Viterbi variable $\omega(s_m)$ is calculated recursively by replacing the sum operation with the max operation in the recursion of $\alpha(s_m)$ (Appendix A-1) as follows:

$$\omega(s_0) = \phi_{s_0,0,p_0} \pi_{0,s_0}, \tag{42}$$

$$\omega(s_m) = \max_{s_{m-1}} \pi_{s_{m-1}s_m} \omega(s_{m-1}) \prod_{j=j_m}^{j_{m+1}-1} \phi_{s_m p_{j-1} p_j}, \tag{43}$$

where an argument s_{m-1} that maximizes the max operation is memorized as $\mathrm{prev}(s_m)$ when calculating $\omega(s_m)$.

In the backward step, the local keys $\mathbf{S}$ are obtained in the reverse order as follows:

$$s_M = \operatorname*{argmax}_i \omega(s_M = i), \tag{44}$$

$$s_m = \mathrm{prev}(s_{m+1}). \tag{45}$$

2) Estimating Musical Notes: In the forward step, a Viterbi variable $\bar{\omega}(\bar{q}_n)$ is calculated recursively by replacing the sum operation with the max operation in the recursion of $\bar{\alpha}(\bar{q}_n)$ (Appendix A-2). In practice, we can introduce weighting factors to balance the musical score model and the F0 trajectory model, as is usually done in statistical speech recognition [38]. The modified message is thus given by

$$\bar{\omega}(\bar{q}_1) = \phi_{s_0,0,\bar{p}_1}^{\beta^{\phi}}, \tag{46}$$

$$\bar{\omega}(\bar{q}_n) = \epsilon_{\bar{e}_n}^{\beta^{\epsilon}} \delta_{\bar{d}_n}^{\beta^{\delta}} \max_{\bar{q}_{\bar{o}_n}} \lambda_{l_{\bar{o}_n}, l_n}^{\beta^{\lambda}} \phi_{s_{\bar{o}_n}, \bar{p}_{\bar{o}_n}, \bar{p}_n}^{\beta^{\phi}} \chi(\bar{q}_n)^{\beta^{\chi}} \bar{\omega}(\bar{q}_{\bar{o}_n}), \tag{47}$$

where an argument $\bar{q}_{\bar{o}_n}$ that maximizes the max operation is memorized as $\mathrm{prev}(\bar{q}_n)$ when calculating $\bar{\omega}(\bar{q}_n)$, and β^{ϕ}, β^{λ}, β^{ϵ}, β^{δ}, and β^{χ} are weighting factors.

Our preliminary experiments show that the latent variables estimated with an HSMM favor longer durations for reducing the number of state transitions because the accumulated multiplication of transition probabilities reduces the likelihood. As a possible solution for penalizing longer musical notes, we introduce an additional term $f(\bar{o}_n) = \{\exp(\frac{1}{n-\bar{o}_n})\}^{\beta^o}$ to Eq. (47) as follows:

$$\bar{\omega}(\bar{q}_n) = \epsilon_{\bar{e}_n}^{\beta^{\epsilon}} \delta_{\bar{d}_n}^{\beta^{\delta}} \max_{\bar{q}_{\bar{o}_n}} \lambda_{l_{\bar{o}_n}, l_n}^{\beta^{\lambda}} \phi_{s_{\bar{o}_n}, \bar{p}_{\bar{o}_n}, \bar{p}_n}^{\beta^{\phi}} \chi(\bar{q}_n)^{\beta^{\chi}} \bar{\omega}(\bar{q}_{\bar{o}_n}) f(\bar{o}_n), \tag{48}$$

where $\bar{o}_n$ and n indicate the onset and offset score times (*i.e.*, $n - \bar{o}_n$ indicates the note value) and β^o is a weighting factor.

In the backward step, the musical notes $\mathbf{Q}$ are obtained in the reverse order as follows:

$$\bar{q}_N = \operatorname*{argmax}_{\bar{q}} \omega(\bar{q}_N = \bar{q}), \tag{49}$$

$$\bar{q}_{\bar{o}_n} = \mathrm{prev}(\bar{q}_n). \tag{50}$$

3) Estimating Model Parameters: Given the latent variables $\mathbf{S}$ and $\mathbf{Q}$, the model parameters $\boldsymbol{\Theta}$ except for σ are obtained as the expectations of the posterior Dirichlet distributions given in Appendix A-3. The Cauchy scale σ is updated to a proposal given by Eq. (39) only when the posterior given by the product of Eqs. (12)–(18) is increased.

REFERENCES

[1] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.

[2] Y. Ojima, E. Nakamura, K. Itoyama, and K. Yoshii, "A hierarchical bayesian model of chords, pitches, and spectrograms for multipitch analysis," in *International Society for Music Information Retrieval Conference, (ISMIR)*, 2016, pp. 309–315.

[3] A. McLeod, R. Schramm, M. Steedman, and E. Benetos, "Automatic transcription of polyphonic vocal music," *Applied Sciences*, vol. 7, no. 12, 2017.

[4] S. Sigtia, E. Benetos, N. Boulanger-Lewandowski, T. Weyde, A. S. d. Garcez, and S. Dixon, "A hybrid recurrent neural network for music transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2061–2065.

[5] A. Ycart and E. Benetos, "A study on LSTM networks for polyphonic music sequence modelling," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 421–427.

[6] E. Nakamura, R. Nishikimi, S. Dixon, and K. Yoshii, "Probabilistic sequential patterns for singing transcription," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1905–1912.

[7] R. Nishikimi, E. Nakamura, M. Goto, K. Itoyama, and K. Yoshii, "Scale- and rhythm-aware musical note estimation for vocal F0 trajectories based on a semi-tatum-synchronous hierarchical hidden semi-markov model," in *International Society for Music Information Retrieval Conference, (ISMIR)*, 2017, pp. 376–382.
[8] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: a new Python audio and music signal processing library," in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 1174–1178.
[9] C. Raphael, "A hybrid graphical model for rhythmic parsing," *Artificial Intelligence*, vol. 137, no. 1-2, pp. 217–238, 2002.
[10] M. Hamanaka, M. Goto, H. Asoh, and N. Otsu, "A learning-based quantization: Unsupervised estimation of the model parameters," in *International Conference on Multimodal Interfaces (ICMI)*, 2003, pp. 369–372.
[11] B. Shahriari, K. Swersky, and Z. Wang, "Taking the human out of the loop: A review of Bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016.
[12] D. J. Hermes, "Measurement of pitch by subharmonic summation," *The journal of the acoustical society of America*, vol. 83, no. 1, pp. 257–264, 1988.
[13] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent F0 estimation and source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 574–578.
[14] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.
[15] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.
[16] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.
[17] M. Mauch and S. Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663.
[18] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 57–60.
[19] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano, "Songle: A web service for active music listening improved by user contributions," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2011, pp. 311–316.
[20] R. P. Paiva, T. Mendes, and A. Cardoso, "On the detection of melody notes in polyphonic audio," in *International Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 175–182.
[21] C. Raphael, "A graphical model for recognizing sung melodies," in *International Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 658–663.
[22] A. Laaksonen, "Automatic melody transcription based on chord transcription," in *International Society for Music Information Retrieval (ISMIR)*, 2014, pp. 119–124.
[23] M. P. Ryynänen and A. P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, 2008.
[24] E. Molina, L. J. Tardón, A. M. Barbancho, and I. Barbancho, "SiPTH: Singing transcription based on hysteresis defined on the pitch-time curve," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 2, pp. 252–263, 2015.
[25] L. Yang, A. Maezawa, J. B. L. Smith, and E. Chew, "Probabilistic transcription of sung melody using a pitch dynamic model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 301–305.
[26] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided melody note transcription using the Tony software: Accuracy and efficiency," in *Proc. of the 1st International Conference on Technologies for Music Notation and Representation (TENOR)*, 2015, pp. 23–30.
[27] H. Takeda, N. Saito, T. Otsuki, M. Nakai, H. Shimodaira, and S. Sagayama, "Hidden markov model for automatic transcription of MIDI signals," in *IEEE Workshop on Multimedia Signal Processing (MMSP)*, 2002, pp. 428–431.
[28] M. Tanji, D. Ando, and H. Iba, "Improving metrical grammar with grammar expansion," in *Artificial Intelligence*, 2008, pp. 180–191.
[29] P. Desain and H. Honing, "The quantization of musical time: A connectionist approach," *Computer Music Journal*, vol. 13, no. 3, pp. 56–66, 1989.
[30] E. Nakamura, K. Yoshii, and S. Sagayama, "Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 794–806, 2017.
[31] Y. W. Teh, "A hierarchical bayesian language model based on Pitman-Yor processes," in *International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006, pp. 985–992.
[32] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.
[33] M. Goto, "AIST annotation for the RWC music database," in *International Conference on Music Information Retrieval (ISMIR)*, 2006, pp. 359–360.
[34] T. De Clercq and D. Temperley, "A corpus analysis of rock harmony," *Popular Music*, vol. 30, no. 01, pp. 47–70, 2011.
[35] E. Molina, A. M. Barbancho, L. J. Tardón, and I. Barbancho, "Evaluation framework for automatic singing transcription," in *International Society for Music Information Retrieval (ISMIR)*, 2014, pp. 567–572.
[36] R. Nishikimi, E. Nakamura, K. Itoyama, and K. Yoshii, "Musical note estimation for F0 trajectories of singing voices based on a bayesian semi-beat-synchronous HMM," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2016, pp. 461–467.
[37] S.-Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, 2010.
[38] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," in *European Conference on Speech Communication and Technology (EUROSPEECH)*, 2001, pp. 1691–1694.

Ryo Nishikimi received the B.E. and M.S. degrees from Kyoto University, Kyoto, Japan, in 2016 and 2018, respectively. He is currently working toward the Ph.D. degree in Kyoto University. His research interests include music informatics and machine learning. He is a member of IEEE and IPSJ.

Eita Nakamura received a Ph.D. degree in physics from the University of Tokyo in 2012. He has been a post-doctoral researcher at the National Institute of Informatics, Meiji University, and Kyoto University. He is currently an Assistant Professor at the Hakubi Center for Advanced Research and Graduate School of Informatics, Kyoto University. His research interests include music modelling and analysis, music information processing and statistical machine learning.

Masataka Goto received the Doctor of Engineering degree from Waseda University in 1998. He is currently a Prime Senior Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. Over the past 28 years he has published more than 270 papers in refereed journals and international conferences and has received 51 awards, including several best paper awards, best presentation awards, the Tenth Japan Academy Medal, and the Tenth JSPS PRIZE. In 2016, as the Research Director he began OngaACCEL Project, a 5-year JST-funded research project (ACCEL) on music technologies.

Katsutoshi Itoyama received the B.E. degree in engineering in 2006, the M.S. degree in informatics in 2008, and the Ph.D. degree in informatics in 2011 all from Kyoto University. He worked with Kyoto University from 2011 to 2018. After that, he is currently an Specially Appointed Associate Professor (Lecturer), Tokyo Institute of Technology. His research interests include musical information processing, robot audition, and computational auditory scene analysis based on statistical signal processing and machine learning techniques. He is a member of the IPSJ, ASJ, and IEEE.

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