SCALE- AND RHYTHM-AWARE MUSICAL NOTE ESTIMATION FOR VOCAL F0 TRAJECTORIES BASED ON A SEMI-TATUM-SYNCHRONOUS HIERARCHICAL HIDDEN SEMI-MARKOV MODEL

First Author
Affiliation1
author1@ismir.edu

Second Author
Affiliation2
Retain these fake authors in submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

ABSTRACT
This paper presents a statistical method that estimates a musically-natural sequence of musical notes from a vocal F0 trajectory. Since the onset times and F0s of sung notes are considerably deviated from the tatum and pitches indicated in a musical score, a score model is crucial for improving time-frequency quantization of F0s. We thus propose a hierarchical hidden semi-Markov model (HHSMM) that combines a score model representing the rhythms and pitches of musical notes under musical scales with an F0 model representing the time-frequency deviations of F0s from the score. In the score model, musical scales are generated stochastically and note pitches are then generated according to the scales. Additionally, note onsets following a Markov process defined on the tatum grid are generated. In the F0 model, onset temporal deviations, smooth note-to-note F0 transitions, and F0 fluctuations are generated stochastically and added to the score. Given an F0 trajectory, our method estimates the most likely sequence of musical notes while giving more importance on the score model than the F0 model. Experimental results showed that the proposed method outperformed an HMM-based method having no models of scales and rhythms.

1. INTRODUCTION
Singing voice analysis is important for music information retrieval because a singing voice usually forms a large part of the melody line of popular music, and provides much information about music. Singing voice analysis techniques such as vocal F0 estimation [1,3,7,9,14] and singing voice separation [8,12] have actively been studied and applied to singer identification [10,22], karaoke generation [19], query-by-humming [8], and active music listening [6]. To leverage musical information conveyed by singing voices, it is necessary to convert a vocal F0 trajectory to a musical score containing only discrete symbols.

In this study, we tackle musical note estimation for a singing voice. This problem is challenging because a vocal F0 trajectory has considerable deviations from the corresponding musical score. The pitches and onset times of musical notes in a musical score can take only discrete values, whereas a vocal F0 trajectory is a continuous signal that can dynamically and smoothly vary over time. For example, the vocal F0 trajectory is modulated by vibrato and changes smoothly from one musical note to another by a portamento. These time-frequency deviations cause a standard transcription system to yield a sequence of musical notes that are inconsistent with music theory.

To solve this problem, we propose a statistical method of musical note estimation based on a hierarchical hidden semi-Markov model (HHSMM) that combines a score model for the process generating a sequence of musical notes and an F0 model for the process generating an F0 trajectory from the note sequence (Fig. 1). A key advantage of the proposed method is that a sequence of musical notes that sound more natural in terms of music theory can be obtained by using musical scales and rhythms as constraints on quantization of vocal F0 trajectories. In the score model, the semitone-level pitch of each note is determined depending on a scale that the note belongs to and the pitch of a previous note. On the other hand, the onset position of a note depends on that of a previous note so that the sequence of notes has appropriate rhythms. Then, the time-frequency fluctuations are added to a step-function-like F0 trajectory exactly corresponding to a musical score. Given a vocal F0 trajectory and tatum information, the scales, musical notes, and deviations, which are latent variables of the proposed model, are jointly estimated by using a Markov chain Monte Carlo algorithm.

![Figure 1](image-url)
2. RELATED WORK

In this section, we introduce related work on the analysis of singing voices.

2.1 Vocal F0 Estimation for Music Audio Signals

Estimation of vocal F0 trajectories for music audio signals has actively been studied [1, 3, 7, 9, 14], and the outputs of these methods can be used as inputs of our method. One of the most basic method is subharmonic summation (SHS) [7] that calculates the sum of the harmonic components of each candidate F0. Ikemia et al. [9] improved F0 estimation based on SHS and singing voice separation based on robust principle component analysis (RPCA) [8] by using the mutual dependency of those two tasks. Salmon et al. [21] estimated contours of the melody F0 candidates by calculating a salience function and then recursively removed contours which do not form a melody line by using the characteristics of each contour. Durrieu et al. [3] extracted a main melody by representing accompaniments with a model inspired by non-negative matrix factorization (NMF) and leading voices with a source-filter model. Mauch et al. [14] modified the YIN [1] in a probabilistic way so that the modified system could determine multiple candidate fundamental frequencies and then choose one at each frame by using an HMM.

2.2 Musical Note Estimation for Singing Voices

Estimation of musical notes from sung melody have been a hot research topic [6, 11, 13, 15, 17, 18, 20, 23]. A naive method is to take the majority of vocal F0s in each interval of a regular grid [6]. Paiva et al. [17] proposed a step-by-step method with five stages: multipitch detection, multipitch trajectory construction, segmentation of multipitch trajectory, elimination of irrelevant notes, and extraction of notes that form a main melody. Raphael [18] proposed an HMM-based method that estimates pitches, rhythms, and tempos when the number of notes is given. The rhythm and onset deviation models used in [18] are similar to those in [15] focused on the hysteresis characteristics of vocal F0s. Nishikimi et al. [16] proposed a method based on an HMM that represents the generative process of a vocal F0 trajectory considering the time and frequency deviations. Yang et al. [23] proposed a method based on a hierarchical HMM that represents the generative process of the \( f_0 - \Delta f_0 \) plane. Mauch et al. [13] developed a software tool called Tony for extracting pitches. In this tool, a vocal F0 trajectory is estimated by PYIN [14], and musical notes are estimated by a modified version of Ryynänen’s method [20].

3. PROPOSED METHOD

This section explains the proposed method of estimating a sequence of musical notes from a vocal F0 trajectory. The method is based on an HHSSM (Fig. 1) that stochastically generates the F0 trajectory with time-frequency deviations from a sequence of musical notes depending on musical scales. The upper part of the proposed model is an HMM that stochastically generates a sequence of musical notes according to the scales of keys that are assigned to bars. The lower part is an HSSM that represents the musical notes and temporal deviations as latent variables and the frequency deviations as F0 emission probabilities.

3.1 Problem Specification

The problem we tackle is defined as follows:

**Input**: A vocal F0 trajectory \( X = \{x_t\}_{t=1}^T \) and 16th-note-level tautums \( Y = \{(u_n, v_n)\}_{n=0}^N \).

**Output**: A sequence of notes \( Z = \{z_j = (p_j, l_j)\}_{j=0}^J \).

where \( T \) is the number of frames in a vocal F0 trajectory, \( x_t \) is a log frequency at time \( t \), and \( N \) is the number of 16th-note-level tautums. \( u_n \in \{1, \ldots, T+1\} \) is the time of tatum \( n \) and the beginning and end of music are represented as \( u_0 = 1 \) and \( u_N = T+1 \), respectively. \( v_n \in \{0, \ldots, 15\} \) is the relative position of tatum \( n \) in a bar. \( J \) is the number of musical notes estimated by proposed methods, and the \( j \)-th note \( z_j \) is represented as a pair consisting of an pitch \( p_j \in \{1, \ldots, K\} \) and a note length \( l_j \in \{1, \ldots, L\} \) in the unit of tautums, where \( K \) is the number of kinds of semitone-level pitches, and \( p_j \) indicates any one in \( \{\mu_1, \ldots, \mu_K\} \), which is a set of log frequencies corresponding to semitone-level pitches. For convenience we introduce the initial note \( z_0 \) that does not appear in the actual score.

3.2 Probabilistic Modeling of Musical Scores

This section describes the score model representing rhythms and pitches of musical notes under musical scales.

3.2.1 Modeling Key Transitions

Keys are represented as \( S = \{s_m\}_{m=0}^M \), where \( M \) denotes the number of bars in the musical piece and \( s_m \) denotes the key at the \( m \)-th bar. For convenience, we introduce the initial bar \( s_0 \) to which the initial note \( z_0 \) belongs. Instead of fixing one key for the whole piece, the key is allowed to change at bar lines. Each key \( s_m \) takes one of the 24 values of \( \{C, C#, \ldots, B\} \times \{\text{major}, \text{minor}\} \). The latent variables \( S \) are described by a Markov chain as

\[
p(s_0 | \pi) = \pi_{s_0},
\]

\[
p(s_m | s_{m-1}, \xi_{s_{m-1}}) = \xi_{s_{m-1} \rightarrow s_m},
\]

where \( \pi \in \mathbb{R}_{\geq 0}^{24} \) is a set of initial probabilities and \( \xi_{s} \in \mathbb{R}_{\geq 0}^{24} \) is a set of transition probabilities.

3.2.2 Modeling Pitch Transitions

The sequence of pitches \( P \) is generated by a Markov chain depending on keys \( S \) as follows (Fig. 2):

\[
p(p_0 | s_0, \phi_{s_0}) = \phi_{s_0 p_0},
\]

\[
p(p_j | p_{j-1}, s_m, \psi_{s_m p_{j-1}}) = \psi_{s_m p_j \rightarrow p_{j-1}}.
\]
where $\phi_r, \psi_{s, p} \in \mathbb{R}_+^K$ is a set of initial probabilities, $\psi_{s, p} \in \mathbb{R}_+^K$ is a set of transition probabilities, and $m$ is the index of a bar to which the note $z_j$ belongs. Moreover, $\phi_{s_0 p_0}$ and $\psi_{s_m p_{j-1} p_j}$ are defined as

$$\phi_{s_0 p_0} = \frac{\phi_{s_0} \deg(p; s_0)}{\sum_{p=1}^K \phi_s \deg(p; s_0)},$$

$$\psi_{s_m p_{j-1} p_j} = \frac{\psi_{s_m} \deg(p_{j-1}; s_m) \deg(p; s_m)}{\sum_{p=1}^K \psi_s \deg(p_{j-1}; s_m) \deg(p; s_m)},$$

where $s \in \{\text{major}, \text{minor}\}$ is the mode of key $s$ and $\deg(p; s) \in \{0, \ldots, 11\}$ is the degree of pitch $p$ in key $s$ (defined as the relative pitch class of $p$ from the tonic of key $s$). $\phi_s$ and $\psi_s$ are the initial and transition probabilities of pitch classes, given the scales.

### 3.2.3 Modeling Onset Transitions

Consider the transition between onset positions of adjacent notes; the model makes $Z$ have the plausible rhythm. Let $r_{j-1} \in \{v_n\}_{n=1}^N$ be the onset position of the $j$-th note $z_j$. The transition probability is given by

$$p(r_j | r_{j-1}, \zeta_{r_{j-1} r_j}) = \zeta_{r_{j-1} r_j},$$

where the distance between $r_{j-1}$ and $r_j$ indicates the note value $l_j$ of note $z_j$. We assume that $r_0 = v_0$ and $r_j = v_N$.

### 3.3 Probabilistic Modeling of F0 Trajectories

The section describes the F0 model that represents the time-frequency deviations of F0s from the musical score.

#### 3.3.1 Modeling Temporal Deviations

We assume that vocal F0 trajectories include the following two types of temporal deviations (Fig. 3a):

**Onset deviation:** the gap between the vocal onset time and the note onset time.

**F0 transitional duration:** the time it takes for singing voices to finish transitioning from one pitch to the next. The onset deviations $G = \{g_j\}_{j=1}^J$ accompanying with $Z$ are represented as discrete latent variables. Each $g_j$ can take an integer value between $-G$ and $G$. As with the onset position model, $g_{j-1}$ denotes the onset deviation of note $z_j$. We assume that each $g_j$ is independently generated by

$$p(g_j | \rho) = \rho g_j,$$

where $\rho \in \mathbb{R}_+^{2G+1}$ is a set of onset deviation probabilities. We assume that there are no deviations for the onset of the first note and the offset of the last note, i.e., $g_0 = g_J = 0$.

### 3.3.2 Modeling Frequency Deviations

The F0 transitional durations $D = \{d_j\}_{j=1}^J$ accompanying with $Z$ are also represented as discrete latent variables. Each $d_j$ can take a value from 1 to $D$. The continuous transition of vocal F0s between notes $z_{j-1}$ and $z_j$ is represented by a slanted line spanning $d_j$ frames. We assume that each $d_j$ is independently generated as follows:

$$p(d_j | \eta) = \eta d_j,$$

where $\eta \in \mathbb{R}_+^D$ is a set of duration probabilities.

### 3.3.3 Modeling Frequency Deviations

The vocal F0 trajectory $X$ is generated by imparting probabilistic frequency deviations to the sequence of notes to which temporal deviations have already been imparted (Fig. 3b). Assuming that $x_t$ is independently generated at each frame, the emission probability of the $j$-th note $z_j$ is given by

$$p(x_{\tau_{j-1} : \tau_j} | p_j, p_j, l_j, g_{j-1}, g_j, d_j, \mu_t, \lambda) = \prod_{t=\tau_{j-1}+1}^{\tau_j} \{\delta_{x_t, \text{voiced}} \text{Cauchy}(x_t | \mu_t, \lambda) + \delta_{x_t, \text{unvoiced}}\} = e_{p_{j-1} p_j l_j g_{j-1} g_j d_j},$$

where $X_{\tau_{j-1} : \tau_j}$ indicates $x_{\tau_{j-1}}, \ldots, x_{\tau_j}$, $\lambda$ is a scale parameter, and $\mu_t$ (Fig. 4) is a location parameter given by

$$\mu_t = \begin{cases} \frac{\mu_{t-1} - \mu_{p_{j-1} - 1}}{d_j} (t - \tau_{j-1}) + \mu_{p_{j-1} - 1} (\tau_{j-1} \leq t \leq \tau_j + d_j) \\ \mu_{g_j} (\tau_{j-1} + d_j \leq t \leq \tau_j) \end{cases}.$$

When the onset of note $z_{j+1}$ is located at the $n$-th tatum, $\tau_j = u_n + g_j$ and $\tau_{j-1} = u_{n-1} + g_{j-1}$.

### 3.4 Prior Distributions

We put conjugate Dirichlet priors on categorical model parameters $\pi, \xi, \phi, \psi, \zeta, \rho$, and $\eta$ as follows:

$$\pi_s \sim \text{Dirichlet}(\alpha_s^T), \ \xi_{sk} \sim \text{Dirichlet}(\alpha_{sk}^T),$$
\[ \phi_s \sim \text{Dirichlet} \left( a_s^\phi \right), \quad \psi_{j_{\text{deg}}(p,s)} \sim \text{Dirichlet} \left( a_{j_{\text{deg}}(p,s)}^\psi \right), \]
\[ \zeta_r \sim \text{Dirichlet} \left( a_r^\zeta \right), \quad \rho \sim \text{Dirichlet} \left( a_r^\rho \right), \]
\[ \eta \sim \text{Dirichlet} \left( a_r^\eta \right) \]

where \( a_s^\phi, a_r^\zeta, a_r^\rho, a_r^\eta \) are hyperparameters.

Since the Cauchy distribution does not have a conjugate prior, we put a Gamma prior on \( \lambda \) as
\[ \lambda \sim \text{Gamma} \left( a_\lambda^0, a_\lambda^1 \right), \]
where \( a_\lambda^0 \) and \( a_\lambda^1 \) are shape and rate hyperparameters.

### 3.5 Bayesian Inference

Given an F0 trajectory \( X \), we aim to calculate the posterior distribution \( p(Q, S, \Theta | X) \), where \( Q = \{ P, L, G, D \} \) (latent variables) and \( \Theta = \{ \pi, \xi, \phi, \psi, \zeta, \rho, \eta \} \) (model parameters).

To get samples of the latent variables \( S \) and \( Q \), forward filtering-backward sampling algorithms are used. To get samples of \( \Theta \) except for \( \lambda \), a set of parameters with conjugate priors, a Gibbs sampling algorithm is used. Since there is no conjugate prior for the parameter \( \lambda \), we use the Metropolis-Hastings (MH) algorithm.

Since \( S \) and \( Q \) share the sequence of notes \( Z \) and are mutually dependent, each variable is updated as follows:

1. Initialize notes \( Z \) with a majority-vote method.
2. Update the sequence of keys \( S \) based on given \( Z \).
3. Update \( Q \) based on given \( S \).
4. Update the model parameters \( \Theta \).
5. Return to 2.

#### 3.5.1 Inferring Latent Variables \( S \)

Given the sequence of notes \( Z \), each \( s_m \) is sampled in accordance with the probability given by
\[ \beta^S_{s_m} = p(s_m | s_{m+1:M}, Z), \]
where \( s_{m+1:M} \) represents \( s_{m+1}, \ldots, s_M \). The calculation of Eq. (14) and sampling of keys \( S \) are performed by the forward filtering-backward sampling method.

In forward filtering, we recursively calculate the probability \( \alpha^S_{s_m} \) as follows:
\[ \alpha^S_{s_0} = p(p_0, s_0) = p(p_0 | s_0) p(s_0) = \phi_{s_0} p_0 \pi_{s_0}, \]
\[ \alpha^S_{s_m} = \prod_{j=m+1}^M \psi_{s_{j-1}} \xi_{s_{j+1} s_{j+1}} \alpha^S_{s_{j+1}}, \]
where \( j_m \) is the index of the first note whose onset belongs to the \( m \)-th bar. \( j_m \) can be calculated from given note values \( L \).

In backward sampling, Eq. (14) is calculated by using the values calculated in forward filtering, and keys are sampled recursively as follows:
\[ \beta^S_{s_M} = p(s_M | Z) \propto \alpha^S_{s_M}, \]
\[ \beta^S_{s_m} = p(s_m | s_{m+1:M}, Z) \propto \alpha^S_{s_m} \xi_{s_m} \psi_{s_{m+1}}, \]

#### 3.5.2 Inferring Latent Variables \( Q \)

The latent variables \( Q \) can be estimated in a way similar to that in which the latent variables \( S \) are inferred. In forward filtering, we recursively calculate the probability
\[ \alpha^Q_{p_n l_n g_n d_n} = p(p_0 | S) = \phi_{p_0}, \]
\[ \alpha^Q_{p_n l_n g_n d_n} = p(x_1, \tau_n = \l_n, g_n, d_n | S) \]

where \( \tau_n = u_n + g_n, \tau'_n = n - l_n \), and \( m(n') \) is the index of the bar that the \( n'-th \) tatum belongs to. \( p_n, l_n, g_n, \) and \( d_n \) are the variables of forward messages that correspond to the note whose offset position is at the \( n \)-th tatum \( u_n \).

Note that these variables are different from \( j \)-indexed variables \( p_j, l_j, g_j, \) and \( d_j \). Since the onset and offset times of the note \( \tau_n = (p_n, l_n) \) are respectively the \( (n-l_n) \)-th tatum and the \( n \)-th tatum, the probability \( p(l_n) \) which appears in the recursive calculation of Eq. (20) is replaced by \( p'(r_n | r_n-l_n) \).

In backward sampling, the posterior distribution of the latent variables is calculated by using the values calculated in forward filtering, and notes and temporal deviations are sampled recursively as follows:
\[ \beta_{p_n l'_n g_n d_n} = p(p_n | l_n, g_n, d_n | X, S) \propto \alpha^Q_{p_n l'_n g_n d_n} \]
\[ \beta_{p_n l'_n g_n d_n} = p(p_n, l_n', g_n, d_n' | p_n, l'_n, g_n, d_n, n, X) \]

#### 3.5.3 Learning Model Parameters \( \Theta \)

The posterior distributions of the model parameters with the prior distributions are calculated using \( S \) and \( Q \) obtained in the backward sampling steps, and these parameters are sampled according to the posterior distributions as follows:
\[ \pi \sim \text{Dirichlet} \left( a^\pi + b^\pi \right), \]
\[ \xi_s \sim \text{Dirichlet} \left( a^\xi_y + b^\xi_y \right), \]
\[ \phi_s \sim \text{Dirichlet} \left( a^\phi_y + b^\phi_y \right), \]
\[ \psi_{j_{\text{deg}}(p,s)} \sim \text{Dirichlet} \left( a^\psi_{\text{deg}(p,s)} + b^\psi_{\text{deg}(p,s)} \right), \]
\[ \zeta_r \sim \text{Dirichlet} \left( a^\zeta + b^\zeta \right), \]
\[ \rho \sim \text{Dirichlet} \left( a^\rho + b^\rho \right), \]
\[ \eta \sim \text{Dirichlet} \left( a^\eta + b^\eta \right), \]
where \( b^* \in \mathbb{R}_{\geq 0}^{26} \) is a unit vector whose \( s_0 \)-th element is 1. \( b^*_r \in \mathbb{R}_{\geq 0}^{26} \) is a vector whose \( s_i \)-th element indicates the number of transitions between adjacent keys \( g \) and \( g' \) in the sequence of latent variables \( Y \). \( b^r \in \mathbb{R}_{\geq 0}^{26} \) is a vector whose \( g \)-th element indicates the number of vocal onset deviations of \( g \) in sampled \( Q \), and \( b^r \in \mathbb{R}_{\geq 0}^{26} \) is a vector whose \( d \)-th element represents the number of F0 transitional durations of \( d \) in sampled \( Q \). \( b^r_s \in \mathbb{R}_{\geq 0}^{16} \) is a vector whose \( r \)-th element represents the number of transitions between adjacent note onset positions \( r \) and \( r' \) in \( R = \{ r_j \} \subset \mathbb{R} \) from that can be calculated from the note values \( L \) sampled in backward sampling. Regarding the vector \( b^r_s \in \mathbb{R}_{\geq 0}^{16} \), when the key of the initial bar and the pitch of the initial note are \( s_0 = s \) and \( p_0 = p \), the value of the element \( b^r_s \) is 1, and the other elements are 0. Regarding the vector \( b^r_s \in \mathbb{R}_{\geq 0}^{16} \), the value of \( b^r_s \) is increased by one when there is a transition from a pitch \( p \) to a pitch \( p' \) under a key \( s \) in the sampled latent variables.

To apply the MH sampling to the parameter \( \lambda \), we define a random-walk proposal distribution as follows:

\[
p(\lambda^* | \lambda) = \text{Gamma}(\gamma \lambda, \gamma),
\]

where \( \lambda^* \) is a proposal, \( \lambda \) is the current sample, and \( \gamma \) is a hyperparameter. The proposal \( \lambda^* \) is accepted as the next sample according to the probability given by

\[
A(\lambda^*, \lambda) = \min \left\{ \frac{L(\lambda^*) p(\lambda|\lambda^*)}{L(\lambda) p(\lambda^*|\lambda)} \right\},
\]

where

\[
L(\lambda) = \text{Gamma}(\lambda|\phi_0^\lambda, \phi_1^\lambda) \prod_{j=0}^J c_{p_{j-1}p_j, j_1, j_2, d_j},
\]

and \( \{p_j, j_1, j_2, d_j\}_{j=0}^J \) are the values sampled in the backward sampling. The value of \( \lambda \) is updated by \( \lambda^* \) only when the value of \( A(\lambda^*, \lambda) \) is larger than a random number sampled from the uniform distribution \( U(0, 1) \).

### 3.6 Viterbi Decoding

The sequence of latent variables \( S \) and \( Q \) are estimated with the Viterbi algorithm with the model parameters that maximize the joint distribution \( P(X, Q, S, \Theta, \Phi) \) in the learning process. As in the inference of latent variables, we initialize \( Z \) by the majority-vote method, \( S \) is estimated based on \( Z \), and then \( Q \) is estimated depending on the \( S \) estimated in the previous step.

In the Viterbi decoding on keys \( S \), the value \( \omega_S^s \) is recursively calculated as follows:

\[
\omega_S^s = \ln \phi_{s_0, x_0} + \ln \pi_{s_1},
\]

\[
\omega_S^s = \sum_{j=1}^{m-1} \ln \psi_{s_{j-1}s_j, p_{j-1}p_j} + \max_{s_{m-1}} \left\{ \ln \xi_{s_{m-1}s_m} + \omega_S^{s_m} \right\}.
\]

In the recursive calculation of \( \omega_S^S \), the previous state \( s_{m-1} \) that maximizes the value of \( \omega_S^{s_m} \) is memorized as \( c_{s_{m-1}}^S \), and the keys \( S \) are recursively estimated as follows:

\[
s_M = \arg \max_{s_M} \omega_{s_M}^S,
\]

\[
s_{m-1} = c_{s_{m-1}}^S.
\]
The results of music note estimation by the proposed method are shown in Table 1. The proposed method outperformed the majority-vote method and the conventional method in terms of both measures. Comparing the tatum-level matching rates obtained by the four versions of the proposed method, we confirmed that the score model significantly improved the performance of musical note estimation. The use of the onset transition probabilities (rhythm constraints) was found to be more effective than that of the key transition probabilities (scale constraints). Although the tatum-level matching rate obtained by the conventional method (68.8%) was close to that obtained by the proposed method, the note-level matching rate obtained by the proposed method (30.9%) was significantly better than that obtained by the conventional method (14.8%). This is a remarkable advantage of the proposed HHSMM that can directly represent both the pitches and durations (onsets and offsets) of musical notes on symbolic musical scores, not on continuous-time piano rolls.

Examples of estimated musical scores are illustrated in Fig. 5. The proposed method yielded the almost accurate musical score except that some notes were merged. To correctly recognize two adjacent notes with the same pitch, it is necessary to refer to original singing voices or music audio signals. The score estimated without considering the score model, on the other hand, included a lot of wrong notes that were inconsistent with music theory. This result also shows the effectiveness of using the score model as musical constraints on musical note estimation.

5. CONCLUSION

This paper presented a statistical method for musical note estimation from a vocal F0 trajectory. Our method is based on an HHSMM that combines a score model (HMM) representing the generative process of a musical score from musical scales with an F0 model (HSMM) representing the generative process of a vocal F0 trajectory with time-frequency deviation from the musical score. We confirmed that the proposed method can yield more musically-natural sequences of musical notes, which significantly improves the perceived quality of estimated results.

One of the most interesting directions of this research is to use the proposed model as a musically-meaningful prior distribution on a vocal F0 trajectory in vocal F0 estimation for music audio signals. We plan to integrate the proposed “language” model that generates an F0 trajectory from a musical score with an acoustic model that generates a spectrogram from the F0 trajectory in a hierarchical Bayesian manner. This enables us to jointly learn the vocal F0 trajectory and musical score from music audio signals. Joint estimation of beat times and F0s is worth investigating to overcome the problem of estimation-error accumulation in the cascaded estimation approach.

6. REFERENCES


