RHYTHM TRANSCRIPTION OF POLYPHONIC MIDI PERFORMANCES BASED ON A MERGED-OUTPUT HMM FOR MULTIPLE VOICES

Eita Nakamura
Kyoto University
enakamura@sap.ist.i.kyoto-u.ac.jp

Kazuyoshi Yoshii
Kyoto University
yoshii@kuis.kyoto-u.ac.jp

Shigeki Sagayama
Meiji University
sagayama@meiji.ac.jp

ABSTRACT

This paper presents a statistical method of rhythm transcription that estimates the quantised durations (note values) of the musical notes in a polyphonic MIDI performance (*e.g.* piano) signal. Hidden Markov models (HMMs) have been used in rhythm transcription to combine a model for music scores and a model describing the temporal fluctuations in music performances. However, when applied to polyphonic music, conventional HMMs have a problem that they are based on representation of polyphonic scores as linear sequences of chords and thus cannot properly describe the structure of multiple voices. We propose a statistical model in which each voice is described with an HMM and polyphonic performances are described as merged outputs from multiple HMMs, based on the framework of merged-output HMM. We develop a rhythm-transcription algorithm based on this model using an efficient Viterbi algorithm. Evaluation results showed that the proposed model outperformed previously studied HMMs for rhythm transcription of polyrhythmic performances.

1. INTRODUCTION

Music transcription is a fundamental problem in music information processing, requiring the extraction of pitch and rhythm information from music audio signals. There have been many studies on converting a music audio signal into a piano-roll representation based on acoustic modelling of musical sound [1, 2]. To obtain a music score, we must recognise quantised note lengths (or note values) of the musical notes in piano rolls. For this purpose, many studies have been devoted to solving the problem of converting MIDI performances to music scores, which is called rhythm transcription or quantisation [3–12]. In accordance with the general trend, statistical modelling has been gathering attention recently in this field.

Hidden Markov models (HMMs) [13] are the most popular models used in recent studies on rhythm transcription [5–10]. Indeed a monophonic score, when represented as a series of musical notes, can naturally be described with a Markov model. In addition, temporal fluctuations in performances can be described by a continuous-space HMM with a latent variable corresponding to time-varying tempos [10, 14, 15].

When HMMs are used for modelling polyphonic music, we immediately face the problem of score representation. A polyphonic score has multilayer structure, where concurrently sounding notes are grouped into several streams or, in music terminology, *voices* [1]. A conventional way is to represent a polyphonic score as a linear sequence of chords [7]. However, this representation may not retain sequential regularities within voices, such as those in polyrhythmic scores. Furthermore, properties of music performance, like the phenomenon of loose synchrony between voices [17, 18], cannot be captured without explicitly modelling the multiple-voice structure.

The purpose of this paper is to construct a statistical model for rhythm transcription that can describe the multiple-voice structure of polyphonic music scores and performances. We construct a model that describes polyphonic performances as merged outputs from multiple component HMMs, each of which describes the generative process of music scores and performances of one voice. Our model is based on the merged-output HMM [19, 20], which has been developed to describe, in an event-driven manner, symbolic data of polyphonic music. We derive an efficient inference algorithm that can simultaneously separate performed notes into voices and estimate their note values. The proposed model is compared with previously studied HMM-based models by evaluating the accuracy of rhythm transcription for piano performances. A complete model description and extended evaluation results will be presented in our forthcoming paper [23].

The main contribution of this study is the construction of a rhythm-transcription algorithm that can explicitly handle multiple voices with guaranteed optimality. A statistical model with multiple-voice structure based on two-dimensional probabilistic context-free grammar (PCFG) models has been studied [11, 12], but the algorithms developed in those studies had to use provided voice information or a pruning technique that would sacrifice optimality.

2. RELATED WORK

In this section, we review previous HMM-based models for rhythm transcription and discuss the problem of polyphonic extensions.

[1] In this paper, a 'voice' means a unit stream of musical notes that can contain chords.

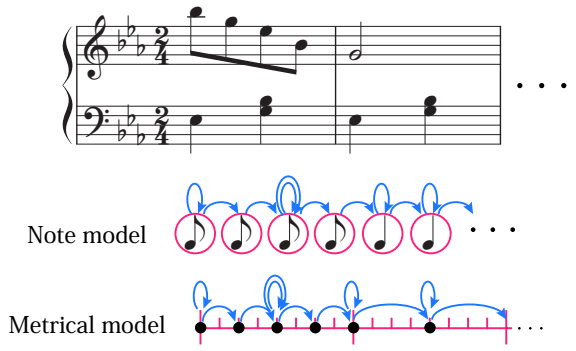

Figure 1. Two different representations of a music score in previously proposed HMMs.

2.1 HMM-Based Models for Monophonic Music

HMMs for rhythm transcription usually consist of two component models; a score model describing the probability of a score and a performance model describing the probability of a performance given a score. HMMs in previous studies [5–10] can be classified into two groups according to the way the score model describes the sequence of notes. In one class of HMMs for rhythm transcription, which we call *note HMMs*, a score is represented as a sequence of note values and described with a Markov model (Fig. 1) [5,6]. To describe the temporal fluctuations in performances, one introduces a latent variable corresponding to a (local) tempo that is also described with a Markov model. An observed duration is described as a product of the note value and the tempo that is exposed to noise of onset times.

In another class of HMMs, which we call *metrical HMMs*, a different description is used for the score model [8–10]. Instead of a Markov model of note values, a Markov process on a grid space representing beat positions of a unit interval, such as a bar, is considered (Fig. 1). The note values are given as differences between successive beat positions. The same performance model as in note HMMs can be used. Incorporation of the metre structure is an advantage of metrical HMMs.

2.2 Polyphonic Extensions

There are two directions of polyphonic extensions: using a simplified representation of polyphonic scores or using an extended model describing multiple voices. The first direction is based on a fact that any polyphonic score can be represented as a sequence of chords or, more precisely, 'note clusters' consisting of one or more notes as far as only onsets are concerned. For note HMMs, chordal notes can be represented as self-transitions in the score model (Fig. 1) and their inter-onset intervals (IOIs) can be described with a probability distribution with a peak at zero [7]. Similar extensions are possible for metrical HMMs.

For the second direction, a PCFG model has been extended to describe the multiple-voice structure of scores [11]. In addition to the divisions of a time interval, duplications of intervals into two voices are considered. Unfortunately, a tractable inference algorithm could not be obtained for this model, and the correct voice information had to be provided for evaluations. Takamune *et al.* state that this problem is solved using the generalised LR parser [12]. Although detailed explanations are lacking, their method uses pruning and its optimality is not guaranteed.

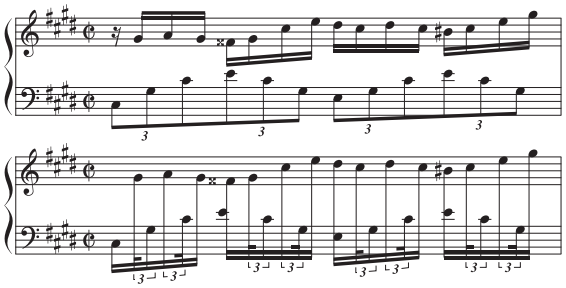

Figure 2. A polyrhythmic passage (Chopin's Fantaisie Impromptu) represented as a sequence of chords.

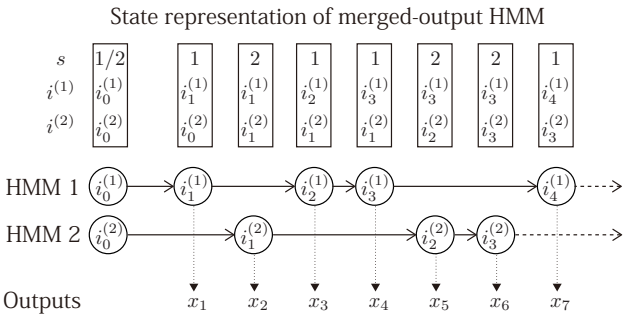

Figure 3. A schematic illustration of the merged-output HMM. The symbols $i_0^{(1)}$ and $i_0^{(2)}$ represent auxiliary states to define the initial transitions.

Although the above two descriptions of polyphonic scores are both logically possible, there are instances in which models based on the simplified representation cannot describe the nature of polyphonic music well. First, complex polyphonic scores such as polyrhythmic scores are forced to have unrealistically small probabilities. This is because such scores consist of rare rhythms in the simplified representation even if the component voices have common rhythms (Fig. 2). Second, the phenomenon of loose synchrony between voices (*e.g.* two hands in piano performances [17]), called *voice asynchrony*, cannot be described. Indeed, the importance of incorporating the multiple-voice structure in describing polyphonic music is well-established in studies on score-performance matching [17, 18]. The situation calls for a similar treatment of multiple voices for polyphonic rhythm transcription.

2.3 Merged-Output HMM

Recently merged-output HMM has been proposed as an HMM-based model for describing symbolic signals of polyphonic music with multiple voices. In the model, each voice is described with an HMM and the total signal is represented as merged outputs from these HMMs (Fig. 3). The merged-output HMM can be seen as a variant of factorial HMM [21]. To appropriately describe the nature of symbolic signals and capture sequential regularities within each voice, only one of the component HMMs is involved with each output in a merged-output HMM, whereas all component HMMs contribute to every output in a standard

factorial HMM. Basic inference algorithms for merged-output HMMs have been provided in our previous studies [19, 20].

3. PROPOSED MODEL

We present an HMM-based model for rhythm transcription that describes polyphonic performances with multiple-voice structure. Given a polyphonic MIDI performance signal, the model can simultaneously separate performed notes into voices and estimate their note values. To construct a model based on a previously studied HMM [7] and apply the framework of merged-output HMM [19, 20], we address the following issues: (1) pitches should be explicitly modelled to appropriately describe voices; (2) tempos of multiple voices should be bound to assure loose synchrony between voices. After explaining the note HMM in detail in Sec. 3.1, a model satisfying these requirements is presented in Sec. 3.2, and a sketch of inference algorithm is given in Sec. 3.3.

A music score is specified by multiple sequences, corresponding to voices, of pitches and note values. Since polyrhythm and voice asynchrony typically involve two voices, we formulate the model with two voices indexed by a variable $s = 1, 2$. A MIDI performance signal is specified by a sequence of pitches and onset times.

3.1 Model for Each Voice

For each voice we first construct a model based on the one presented in a previous study [7]. Let N_s be the number of score notes in voice s and let $r_n^{(s)}$ denote the note value of the n-th note. The note values $\boldsymbol{r}^{(s)} = (r_n^{(s)})_{n=1}^{N_s}$ are generated by a Markov chain with the probability given as

$$r_1^{(s)} \sim \mathsf{Cat}(\boldsymbol{\pi}_{\mathrm{ini}}^{(s)}), \tag{1}$$

$$r_n^{(s)} | r_{n-1}^{(s)} \sim \mathsf{Cat}(\boldsymbol{\pi}_{r_{n-1}^{(s)}}^{(s)}) \quad (n = 2, \ldots, N_s), \tag{2}$$

where Cat denotes the categorical distribution, $\boldsymbol{\pi}_{\mathrm{ini}}^{(s)} = (\pi_{\mathrm{ini},r}^{(s)})_r$ is the initial probability, and $\boldsymbol{\pi}_{r_{n-1}^{(s)}}^{(s)} = (\pi_{r_{n-1}^{(s)},r}^{(s)})_r$ is the (stationary) transition probability. Chordal notes are represented as self-transitions of note values (Fig. 1). The probability values are to be learned from music data.

To describe the temporal fluctuations, we introduce a tempo variable, denoted by $v_n^{(s)}$, that describes the local tempo for the n-th note. To represent the variation of tempos, we put a Gaussian Markov process on the logarithm of the tempo variables as

$$\ln v_n^{(s)} | \ln v_{n-1}^{(s)} \sim \mathsf{N}(\ln v_{n-1}^{(s)}, \sigma_v^2), \tag{3}$$

where N denotes the normal distribution. If the $(n-1)$-th and n-th notes belong to a chord, their IOI approximately obeys an exponential distribution [15] and the probability of the onset time of the n-th note, denoted by $t_n^{(s)}$, is then given as

$$t_n^{(s)} | t_{n-1}^{(s)} \sim \mathsf{Exp}(\lambda), \tag{4}$$

where Exp denotes the exponential distribution and λ is the scale parameter. Otherwise, $t_n^{(s)} - t_{n-1}^{(s)}$ has a duration corresponding to note value $r_{n-1}^{(s)}$ and the probability is described with a normal distribution as

$$t_n^{(s)} | t_{n-1}^{(s)}, v_{n-1}^{(s)}, r_{n-1}^{(s)} \sim \mathsf{N}(t_{n-1}^{(s)} + r_{n-1}^{(s)} v_{n-1}^{(s)}; \sigma_t^2). \tag{5}$$

The measured values of the parameters are $\sigma_t = 0.02$ s and $\lambda = 0.0101$ s [15] (the value of σ_v will be explained later). Remarks should be made here: First, the number of observed onsets must be N_s+1 so that there are N_s IOIs corresponding to N_s score notes. Second, we do not put a distribution on the onset time of the first note $t_1^{(s)}$ because we formulate the model to be invariant under time translations and this value would not affect any results of inference. We will use the notation $\boldsymbol{v}^{(s)} = (v_n^{(s)})_{n=1}^{N_s}$ and $\boldsymbol{t}^{(s)} = (t_n^{(s)})_{n=1}^{N_s+1}$.

Finally we describe the generation of pitches $\boldsymbol{p}^{(s)} = (p_n^{(s)})_{n=0}^{N_s+1}$ as a Markov chain (we introduce an auxiliary symbol $p_0^{(s)}$ for later convenience). The probabilities are

$$p_1^{(s)} | p_0^{(s)} \sim \mathsf{Cat}(\boldsymbol{\theta}_{p_0^{(s)}}^{(s)}), \tag{6}$$

$$p_n^{(s)} | p_{n-1}^{(s)} \sim \mathsf{Cat}(\boldsymbol{\theta}_{p_{n-1}^{(s)}}^{(s)}) \quad (n = 2, \ldots, N_s+1), \tag{7}$$

where $\boldsymbol{\theta}_{p_0^{(s)}}^{(s)} = (\theta_{p_0^{(s)},p}^{(s)})_p$ is the initial probability, and $\boldsymbol{\theta}_{p_{n-1}^{(s)}}^{(s)} = (\theta_{p_{n-1}^{(s)},p}^{(s)})_p$ is the (stationary) transition probability. These parameters are to be learned from music data.

The above model can be summarised as an autoregressive HMM, which we call a voice HMM, with hidden states $(\boldsymbol{r}^{(s)}, \boldsymbol{v}^{(s)})$ and outputs $(\boldsymbol{p}^{(s)}, \boldsymbol{t}^{(s)})$. Although so far the probabilities of pitches are independent of other variables, they will be significant once multiple voice HMMs are merged and the posterior probabilities are inferred.

3.2 Model for Multiple Voices

We combine the multiple voice HMMs in Sec. 3.1 using the framework of merged-output HMMs [19]. Simply speaking, the sequence of merged outputs is obtained by gathering the outputs of the voice HMMs and sorting them according to onset times. To derive inference algorithms that are computationally tractable, however, we should formulate a model that outputs notes incrementally in the order of observations. This can be done by introducing stochastic variables $\boldsymbol{s} = (s_n)_{n=1}^{N+1}$, which indicate that the n-th observed note belongs to voice s_n, with the following probability:

$$s_n \sim \mathsf{Ber}(\alpha_1, \alpha_2), \tag{8}$$

where Ber is the Bernoulli distribution. α_{s_n} represents how likely the n-th note is generated from the HMM of voice s_n and, to improve the results of voice separation, we put on the parameter conditional dependence on the lowest and highest pitches of simultaneously sounding notes.

If voice s_n is chosen, then the HMM of voice s_n outputs a note, and the hidden state of the other voice HMM is unchanged. Such a model can be described with an HMM with a state space labelled by $k_n = (s_n, p_n^{(1)}, r_n^{(1)}, t_n^{(1)}, p_n^{(2)}, r_n^{(2)}, t_n^{(2)}, v_n)$. Here we have a single tempo variable v_n that is shared by the two voices in order to assure loose synchrony between them. $P(k_n | k_{n-1})$,

for $n \geq 2$, is given as

$$\alpha_{s_n} P(v_n|v_{n-1}) A^{(s_n)}_{r^{(s_n)}_{n-1} r^{(s_n)}_n}(p^{(s_n)}_n, t^{(s_n)}_n | p^{(s_n)}_{n-1}, t^{(s_n)}_{n-1}; v_{n-1}) \cdot \Big[\delta_{s_n 1} \delta_{r^{(2)}_{n-1} r^{(2)}_n} \delta_{p^{(2)}_{n-1} p^{(2)}_n} \delta(t^{(2)}_{n-1} - t^{(2)}_n) + (1 \leftrightarrow 2)\Big], \quad (9)$$

where we have defined

$$A^{(s)}_{r^{(s)}_{n-1} r^{(s)}_n}(p^{(s)}_n, t^{(s)}_n | p^{(s)}_{n-1}, t^{(s)}_{n-1}; v_{n-1}) = \pi^{(s)}_{r^{(s)}_{n-1}, r^{(s)}_n} \theta^{(s)}_{p^{(s)}_{n-1}, p^{(s)}_n} P(t^{(s)}_n | t^{(s)}_{n-1}, v_{n-1}, r^{(s)}_{n-1}) \quad (10)$$

and δ denotes Kronecker's delta for discrete variables and Dirac's delta function for continuous variables. The probability $P(v_n|v_{n-1})$ is defined in Eq. (3), and $P(t^{(s_n)}_n | t^{(s_n)}_{n-1}, v_n, r^{(s_n)}_n)$ is defined in Eqs. (4) and (5). For note values the initial probability is given as $r^{(s)}_1 \sim \mathsf{Cat}(\boldsymbol{\pi}^{(s)}_{\mathrm{ini}})$, and for pitches the initial probability is set as in Eq. (6). The first onset times $t^{(1)}_1$ and $t^{(2)}_1$ do not have distributions, as explained in Sec. 3.1, and we practically set $t^{(1)}_1 = t^{(2)}_1 = t_1$ where t_1 is the first observed onset time. Finally the output of the model is given as

$$p_n = p^{(s_n)}_n, \quad t_n = t^{(s_n)}_n, \quad (11)$$

and thus the complete-data probability is written as

$$P(\boldsymbol{k}, \boldsymbol{p}, \boldsymbol{t}) = \prod_n P(k_n|k_{n-1}) \delta_{p_n p^{(s_n)}_n} \delta(t_n - t^{(s_n)}_n). \quad (12)$$

$N = N_1 + N_2$ denotes the total number of score notes, and the following notations will be used: $\boldsymbol{v} = (v_n)^N_{n=1}$, $\boldsymbol{p} = (p_n)^{N+1}_{n=1}$, $\boldsymbol{t} = (t_n)^{N+1}_{n=1}$, and $\boldsymbol{k} = (k_n)^{N+1}_{n=1}$. Note that whereas $\boldsymbol{p}$ and $\boldsymbol{t}$ are observed quantities, $\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}, \boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}$ are not because we cannot directly observe the voice information encrypted in $\boldsymbol{s}$.

3.3 Inference Algorithm

Rhythm transcription based on the proposed model can be performed by estimating the most probable hidden state sequence $\hat{\boldsymbol{k}}$ given the observations $(\boldsymbol{p}, \boldsymbol{t})$. Once $\hat{\boldsymbol{k}}$ is obtained, we can extract the voice information $\hat{\boldsymbol{s}}$ and the note values $\hat{\boldsymbol{r}}^{(1)}$ and $\hat{\boldsymbol{r}}^{(2)}$. These are the result of voice separation and rhythm transcription.

The maximisation of the probability $P(\boldsymbol{k}|\boldsymbol{p}, \boldsymbol{t})$ can be in principle done with the Viterbi algorithm [13]. However, due to the complexity of our model, we need refinements to the standard Viterbi algorithm to derive a computationally tractable algorithm. First, since the state space of the merged-output HMM in Sec. 3.2 involve both discrete and continuous variables, an exact inference is not computationally tractable. To solve this problem, we discretise the tempo variable in a range that is common in music practice. Other continuous variables $\boldsymbol{t}, \boldsymbol{t}^{(1)}$, and $\boldsymbol{t}^{(2)}$ can take only values of observed onset times and thus can, in effect, be treated as discrete variables.

Second, it appears that a Viterbi algorithm derived in the way proposed in [19] has rather large computational cost for the present model and in practice difficult to execute. The large computational cost derives from the fact that we need to model pitches and onset times for the voice HMMs. This problem can be reduced by noting that the pitch and onset time are observed quantities and can be represented by a variable describing the historical information of voices associated to notes, as suggested in [20]. Extending the formalism of introducing a latent variable to describe this information, we can derive an efficient algorithm. Details will be given in our forthcoming paper [23]. We have confirmed that this algorithm can be executed in a standard modern computer environment with a practical time (within a few hours for a performance with hundreds of notes).

4. EVALUATION

4.1 Setup

We evaluated the proposed model by comparing the accuracy of its rhythm transcription with that of previously studied models based on HMMs. Two data sets of MIDI recordings of classical piano pieces were used. One ('polyrhythmic' data set) consisted of 18 performances of 15 (excerpts of) pieces that contained 2 against 3 or 3 against 4 polyrhythmic passages, and the other ('standard polyphony' data set) consisted of 30 performances of 22 pieces that did not contain polyrhythmic passages. Pieces by various composers, ranging from J. S. Bach to Debussy, were chosen and the players were also various: Some of the performances were taken from the PEDB database [22], a few were performances we recorded, and the rest was taken from public domain websites.

All normal, dotted, and triplet note values ranging from the whole note to the 32nd note were used as candidate note values. The transition and initial probabilities of the note values and pitches, and the value of α_s, were learned from a data set of classical piano scores that had no overlap with the test data. For the tempo variable, we discretised v_n into 50 values logarithmically equally spaced in the range of 0.3 to 1.5 sec per quarter note (corresponding to 200 BPM and 40 BPM). The standard deviation in Eq. (3) was set as $\sigma_v = 1.08$, using the value in [15] as a reference.

For comparison, we implemented the note HMM [6] and the metrical HMM [8] that is extended to handle polyphony. The parameters of the score models were also trained with the same score dataset. The performance model was the same as that for the proposed model.

We used as an evaluation measure the rhythm correction ratio, *i.e.*, the ratio of the smallest number of edit operations needed to correct the estimated result to the number of notes in the data. In addition to note-wise correction (shift operation), the scaling operation applied for a subsequence of note values was included. This is because there is arbitrariness in choosing the unit of note values: For example, a quarter note played in a tempo of 60 BPM has the same duration as a half note played in a tempo of 120 BPM. The smallest number of necessary edit operations N_{e} can be calculated by a dynamic programming similar to that used in computation of the Levenshtein distance (see our forthcoming paper [23] for details). The rhythm correction ratio $\mathscr{R}$ is then given as $\mathscr{R} = N_{\mathrm{e}}/N$. When separated voices are given, we can apply the above editing of note values for each voice and then the total rhythm correction cost is the sum of the rhythm correction costs in all voices.

Data set	Model	$\mathscr{R}$ [%]
Polyrhythmic	Proposed	$\mathbf{16.0 \pm 3.6}$
	Note HMM [6]	28.9 ± 4.9
	Metrical HMM [8]	34.1 ± 5.0
Standard polyphony	Proposed	7.9 ± 1.3
	Note HMM [6]	$\mathbf{7.0 \pm 1.3}$
	Metrical HMM [8]	7.9 ± 1.4

Table 1. Average rhythm correction rates $\mathscr{R}$ with standard errors. Lower is better.

4.2 Results

Results in Table 1 show that the proposed model clearly outperformed the other models for performances with polyphonic passages. Fig. 4 shows an example that a polyrhythmic passage is successfully transcribed with the proposed model with minor errors[2]. We see that the proposed model correctly recognised the 3 against 4 polyrhythms. On the contrary, the Note HMM did not recognise the polyrhythms (cf. Fig. 2) and had frequent errors in chord clustering.

For performances in standard polyphony, on the other hand, the note HMM was slightly better than the proposed model and the metrical HMM. Presumably, the main reason is that the rhythmic pattern in the reduced sequence of chords is often simpler than that of melody/chords in each voice in the case of standard polyphony because of the principle of complementary rhythm [24]. In particular, notes/chords in a voice can have tied note values that are not contained in our candidate list (*e.g.* quarter note + 16th note value), which can also appear as a result of incorrect voice separation (Fig. 5). It is also observed that the transcription by the merged-output HMM can produce desynchronised cumulative note values in different voices. This is due to the lack of constraints to assure the matching of these cumulative note values and the simplification of independent voice HMMs. Further improvements are expected by incorporating such constraints and interactions between voices into the model.

For the note HMM and the proposed model, there were grammatically wrong sequences of note values, for example, triplets that appear in single or two notes without completing a unit of beat. This can be avoided with a refined score model with beat/bar structure [6, 11]. On the other hand, these grammatical errors were not observed in the transcriptions by the metrical HMM owing to the explicitly included metrical structure.

5. CONCLUSION

To develop a rhythm transcription algorithm that captures the voice structure, we constructed a stochastic model of musical score and performance using the framework of merged-output HMMs. The evaluation results confirmed that the proposed algorithm worked better for polyrhythmic performances than the previously proposed HMM-based algorithms.

[2] Sound files and more examples are accessible in our demonstration web page: `http://anonymous4721029.github.io/demo.html`

An important future direction of developing advanced transcription techniques is to capture the phrase or motivic structure of music. Recognition of offsets and articulations and detection of ornaments are challenging problems. The treatment of voice structure is a fundamental problem for these issues, and the results of this study may be applicable to solving these problems.

Acknowledgments

This work is partially supported by JSPS KAKENHI Nos. 24220006, 26240025, 26280089, 26700020, 15K16054, 16H01744 and 16J05486, JST OngaCREST Project and Kayamori Foundation. E. Nakamura is supported by the JSPS fellowship program.

6. REFERENCES

[1] A. Klapuri and M. Davy (eds.), *Signal Processing Methods for Music Transcription*, Springer, 2006.

[2] E. Benetos *et al.*, "Automatic Music Transcription: Challenges and Future Directions," *J. Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[3] H. Longuet-Higgins, *Mental Processes: Studies in Cognitive Science*, MIT Press, 1987.

[4] P. Desain and H. Honing, "The Quantization of Musical Time: A Connectionist Approach," *Comp. Mus. J.*, vol. 13, no. 3, pp. 56–66, 1989.

[5] T. Otsuki *et al.*, "Musical Rhythm Recognition Using Hidden Markov Model (in Japanese)," *J. Information Processing Society of Japan*, vol. 43, no. 2, pp. 245–255, 2002.

[6] H. Takeda *et al.*, "Hidden Markov Model for Automatic Transcription of MIDI Signals," *Proc. MMSP*, pp. 428–431, 2002.

[7] H. Takeda *et al.*, "Rhythm and Tempo Analysis Toward Automatic Music Transcription," *Proc. ICASSP*, vol. 4, pp. 1317–1320, 2007.

[8] C. Raphael, "Automated Rhythm Transcription," *Proc. ISMIR*, pp. 99–107, 2001.

[9] M. Hamanaka *et al.*, "A Learning-Based Quantization: Unsupervised Estimation of the Model Parameters," *Proc. ICMC*, pp. 369–372, 2003.

[10] A. Cemgil and B. Kappen, "Monte Carlo Methods for Tempo Tracking and Rhythm Quantization," *J. Artificial Intelligence Res.*, vol. 18 no. 1, pp. 45–81, 2003.

[11] M. Tsuchiya *et al.*, "Probabilistic Model of Two-Dimensional Rhythm Tree Structure Representation for Automatic Transcription of Polyphonic MIDI Signals," *Proc. APSIPA*, pp. 1–6, 2013.

[12] N. Takamune *et al.*, "Automatic Transcription from MIDI Signals of Music Performance Using 2-Dimensional LR Parser (in Japanese)," *Tech. Rep. SIGMUS*, vol. 2014-MUS-104, no. 7, pp. 1–6, 2014.

[13] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

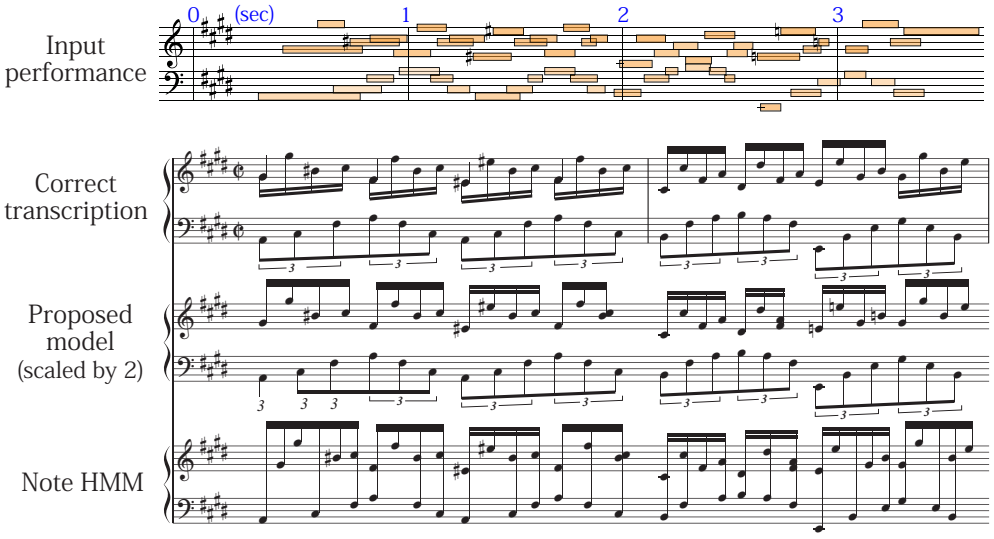

Figure 4. Transcription results of a polyrhythmic passage. For the result with the proposed model (merged-output HMM), the staffs indicate the estimated voices.

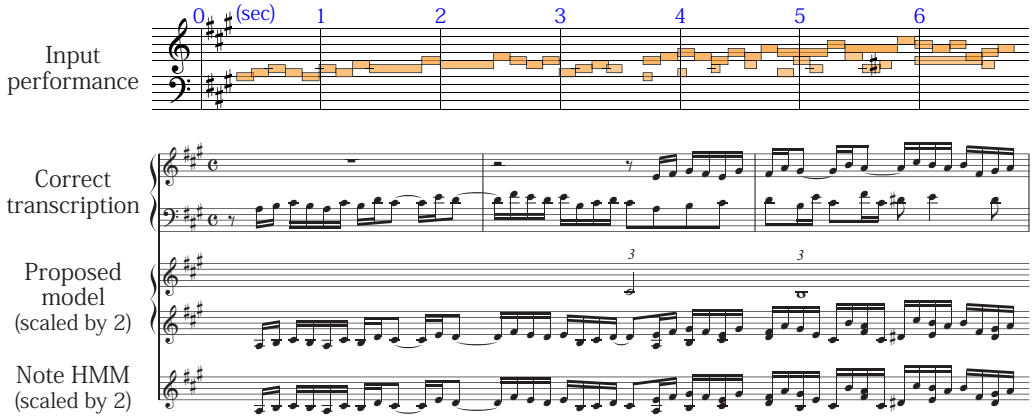

Figure 5. Transcription results of a standard polyphonic passage. For the result with the proposed model (merged-output HMM), the staffs indicate the estimated voices.

[14] C. Raphael, "Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models," *IEEE Trans. on PAMI*, vol. 21, no. 4, pp. 360–370, 1999.

[15] E. Nakamura *et al.*, "A Stochastic Temporal Model of Polyphonic MIDI Performance with Ornaments," *J. New Music Res.*, vol. 44, no. 4, pp. 287–304, 2015.

[16] A. Cont, "A Coupled Duration-Focused Architecture for Realtime Music to Score Alignment," *IEEE Trans. on PAMI*, vol. 32, no. 6, pp. 974–987, 2010.

[17] H. Heijink *et al.*, "Data Processing in Music Performance Research: Using Structural Information to Improve Score-Performance Matching," *Behavior Research Methods, Instruments, & Computers*, vol. 32, no. 4, pp. 546–554, 2000.

[18] B. Gingras and S. McAdams, "Improved Score-Performance Matching Using Both Structural and Temporal Information from MIDI Recordings," *J. New Music Res.*, vol. 40, no. 1, pp. 43–57, 2011.

[19] E. Nakamura *et al.*, "Merged-Output Hidden Markov Model for Score Following of MIDI Performance with Ornaments, Desynchronized Voices, Repeats and Skips," *Proc. Joint ICMC|SMC 2014*, pp. 1185–1192, 2014.

[20] E. Nakamura *et al.*, "Merged-Output HMM for Piano Fingering of Both Hands," *Proc. ISMIR*, pp. 531–536, 2014.

[21] Z. Ghahramani and M. Jordan, "Factorial Hidden Markov Models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

[22] M. Hashida *et al.*, "A New Music Database Describing Deviation Information of Performance Expressions," *Proc. ISMIR*, pp. 489–494, 2008.

[23] E. Nakamura *et al.*, in preparation.

[24] F. Salzer and C. Schachter, *Counterpoint in Composition: The Study of Voice Leading*, Columbia University Press, 1989.