Hidden Markov Model for Automatic Transcription of MIDI Signals

Haruto TAKEDA[1] Naoki SAITO[1] Tomoshi OTSUKI[2]
Mitsuru NAKAI[2] Hiroshi SHIMODAIRA[2] Shigeki SAGAYAMA[1]
[1] Graduate School of Information Science and Technology, The University of Tokyo
Hongo, Bunkyo-ku, Tokyo 113-8656 Japan / {takeda,sagayama}@hil.t.u-tokyo.ac.jp
[2] Graduate School of Information Science, Japan Advanced Institute of Science and Technology
Tatsu-no-kuchi, Ishikawa 923-1292 Japan / {mit,sim}@jaist.ac.jp

Abstract— **This paper describes a Hidden Markov Model (HMM)-based method of automatic transcription of MIDI (Musical Instrument Digital Interface) signals of performed music. The problem is formulated as recognition of a given sequence of fluctuating note durations to find the most likely intended note sequence utilizing the modern continuous speech recognition technique. Combining a stochastic model of deviating note durations and a stochastic grammar representing possible sequences of notes, the maximum likelihood estimate of the note sequence is searched in terms of Viterbi algorithm. The same principle is successfully applied to a joint problem of bar line allocation, time measure recognition, and tempo estimation. Finally, durations of consecutive n notes are combined to form a "rhythm vector" representing tempo-free relative durations of the notes and treated in the same framework. Significant improvements compared with conventional "quantization" techniques are shown.**

I. Introduction

Automatic transcription of music performed on MIDI music instruments has wide applicability including score printing, automatic playing of music pieces, aids for music composition and arrangement, and educational purposes. The problem, however, is not simple even though the pitch of each note is known in the MIDI format; music note durations in human performance fluctuate and intended time values are not easily retrieved from the observation.

The conventional way of treating this problem is *quantization* of observed note durations, music being played synchronously with metronome at a specified tempo [1]. It basically fits fractional note durations to the specified time resolution. This simple method is not applicable to music performances without metronome and changing tempo. Transcribed score far from the intended score is often (almost everytime) experienced among the users. Because of low performance of this method, new quantization models have been investigated [2].

On the other hand, trained humans can easily transcribe performed (relatively simple) music even when the tempo slowly changes. This problem, thus, is considered to essentially involve rhythm pattern recognition utilizing top-down information, while the above previous works took bottom-up approaches.

From this point of view, we previously introduced stachastic modeling based on Hidden Markov Model (HMM) for recognition of the rhythm pattern from given performed music [3] [4] since the rhythm recognition problem is analogous to continuous speech recognition and Hidden Markov Model(HMM)[5] fits both problems. This framework was extended to transcribe music from performance with changing tempo (without metronome), tempo estimation, bar line allocation, and time measure recognition all at the same time. Our approach, first published in Japanese before this English publication, has been already applied to onset time quantization in jam sessions[6].

We also discuss "rhythm vector", a tempo-free rhythm observation feature, in combination with HMM to enable rhythm recognition of performance without estimating the tempo.

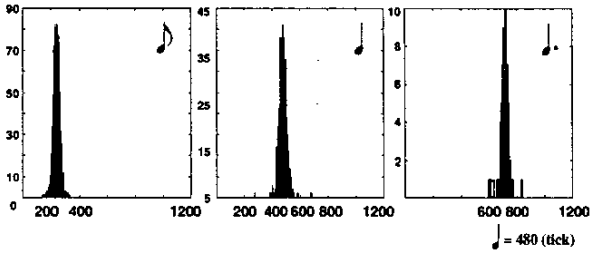

Fig. 1. Distribution of performed note durations.

II. Stochastic Modeling

A. Model of fluctuating note durations

The duration of music notes played by human deviates from the ideal length notated in the score even when the metronome signal is heard. Hereinafter, we call "length" for the ideal (nominal, intended, time value) duration of a note and "duration" for its observed (performed) duration. Fig. 1 shows the distribution of durations of eighth-notes, quarter-notes, and dotted quarter-notes in music pieces performed on a MIDI keyboard by 50 players with a specified tempo (96 by metronome, i.e., one beat = 480 ticks). The note duration is defined as the IOI (inter-onset-time interval).

This figure implies that the fluctuation can be modeled by a Gaussian distribution around the ideal length. When the intention is identified by i (equivalent to the state number in the next subsection) at time t, the performed duration, x_t, is modeled by a probability density function (pdf), $b_i(x_t)$. When the sequence of intentions is identified by a time sequence $Q = \{q_1, q_2, \cdots, q_N\}$, the probability of observing the en-

0-7803-7713-3/02/$17.00 © 2002 IEEE

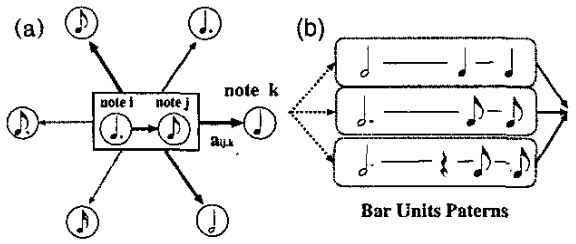

Fig. 2. Note sequence model. (a) 3-gram model, (b) rhythm vocabulary model.

tire sequence $X = \{x_1, x_2, \cdots, x_N\}$ is given by: $P(X|Q) = \prod_{t=1}^{N} b_{q_t}(x_t)$.

B. Model of possible note sequences

When a music is performed, man often can give a reasonable interpretation for the heard sequence of note durations and, as the result, can recognize the intended rhythm pattern. The inference is based on his/her knowledge about possible rhythm patterns acquired through music experiences. This knowledge can be compared to a stochastic language model in modern continuous speech recognition technology.

This aspect is modeled as stochastic generation of intended note length sequences which underlies generally in music depending on genres, styles, and composers. We use two types of rhythm pattern generation models to characterize possible rhythms as follows:

Note n-gram Model: Note length is predicted from preceding $(n-1)$ notes in the probabilistical sense. This model covers any rhythm patterns and can give a certain probability while grammatical constraint is rather weak for small n.

Rhythm Vocabulary Model: The "rhythm vocabulary" consists of all known rhythm patterns for a unit time (typ., one measure). This model well represents known rhythm patterns while unknown patterns are substituted by similar existing patterns.

As shown in Fig. 2, both models are represented by probabilistic state transition networks where each state is associated with an intended note lengths. Labeling all distinct states with integral numbers, $1, 2, \cdots, S$, probability a_{ij} of transition from state i to j characterizes grammatical constraints.

The probability that state number changes along a time sequence $Q = \{q_1, q_2, \cdots, q_N\}$ (q_t: integer) is thus given by $P(Q) = \pi_{q_0} \prod_{t=1}^{N} a_{q_{t-1}q_t}$ where π_i denotes the initial probability of starting the state transition with state i.

We trained model parameters $A = \{a_{ij}\}$ of both rhythm grammar models through statistical estimation. The n-gram model was trained using approximately 50000 notes in MIDI data of classical and jazz music and smoothed by linear combination of probabilities from 1-gram (unigram) through $(n-1)$-gram. The rhythm vocabulary model consisted of 267 one-bar-long rhythm patterns obtained from 88 music pieces including children's songs and folk songs. Connection probabilities between vocabulary words were also obtained from the number of occurrences in the data.

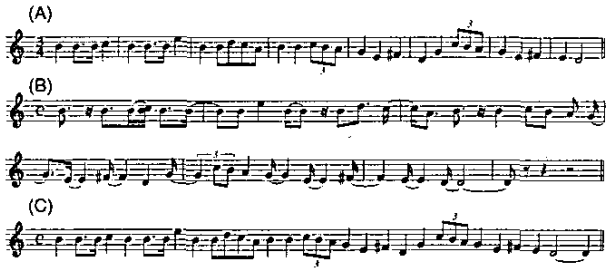

Fig. 3. A typical example of automatic transcription results. A: Testing phrase from Brahms' Symphony No. 1, B: Score obtained by XGworks (quantization), C: Score obtained by n-gram HMM.

C. Model integration by HMM

The models of fluctuating durations and possible note sequences can be combined in the Hidden Markov Model (HMM) framework with transition probabilities, $A = \{a_{ij}\}$, and observation probabilities, $B = \{b_i(x_t)\}$. In a simple bigram (2-gram) model, q_i directly corresponds to the kind of distinct note length. The probability of observing a duration sequence X is given by $P(X|Q)P(Q)$.

III. Rhythm Recognition

A. Inverse problem

Our problem is to find the time sequence of state numbers in the state transition network, Q, that gives the maximum *a posteriori* probability, $P(Q|X)$, given a sequence of observed durations, X. According to the Bayes theorem: $P(Q|X) = \frac{P(X|Q)P(Q)}{P(X)}$, maximizing $P(Q|X)$ is equivalent to finding $\hat{Q} = \underset{Q}{\operatorname{argmax}} P(X|Q)P(Q)$ among all possible Qs. Since the integrated model is represented by an HMM, the optimal sequence of states is efficiently found through the well-known Viterbi algorithm for searching the best path in the probabilistic transition network. The sequence of intended notes is estimated in the maximum likelihood sense as the sequence of notes associated to the states along the best path. This process is referred to rhythm recognition of performed music.

B. Rhythm recognition performance

A typical result of HMM-based rhythm recognition is shown in Fig. 3 and compared with that of quantization by "XGworks" from YAMAHA Corp. when played in a specified tempo. While simple quantization of XGworks inserted numerous wrong rests and ties, HMM almost correctly estimated musical rhythms including triplets. Table I shows the recognition rates of correct note lengths, counting all substitutions, insertions and deletions as errors. "Pause-neglected" recognition scores mean compensated scores ignoring deceptive pauses caused by repeating notes, staccatos, etc.

C. Constant tempo estimation

Unknown constant tempo is estimated in the same framework as stated above. Multiple rhythm-dependent HMMs each representing a different tempo are run to find the maximum likelihood tempo among tempo-dependent models. In our experiments, 6 parallel models were used to represent logarithmically

THE RHYTHM RECOGNITION RATES [%]

method	pauses counted	pauses neglected
Rhythm Vocabulary HMM	59.7	97.3
Bigram HMM	53.7	87.4
Quantization (XGworks)	40.7	85.9

TABLE II

CONSTANT TEMPO ESTIMATION RESULTS [BEATS/MIN.]

Player#	1	2	3	4	5
True Tempo	98.4	93.3	99.2	127.1	106.3
Estimated	95	95	95	120	107
Player#	6	7	8	9	10
True Tempo	116.4	111.7	99.9	109.3	65.2
Estimated	120	107	95	107	67

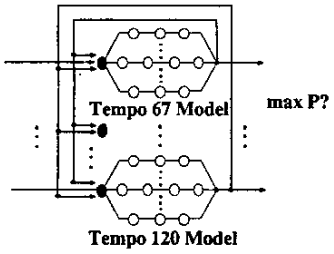

Fig. 4. Model of changing tempo.

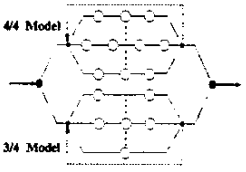

Fig. 5. Model of time measures.

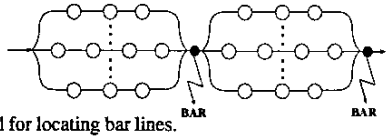

Fig. 6. Model for locating bar lines.

equally spaced tempos between 60 and 120 beats per minute (i.e., 67, 76, 85, 95, 107 and 120 beats/min.) and simultaneously recognized the rhythm Q and tempo T, i.e., maximizing $P(X|Q,T)P(Q|T)P(T)$ in respect to Q and T for the given X. Table II shows a few examples of tempo estimation of performances by 10 players who played a piece shown in Fig. 7.

D. Fluctuating tempo estimation

The same framework with slightly modified models can handle fluctuating tempos. As shown in Fig. 7, models of different tempos are loosely coupled with appropriate probabilities. The maximum likelihood path found through the Viterbi search indicate the recognized rhythm and instantaneous tempos.

One extreme example is shown in Fig. 7 where the fluctuating tempo is successfully detected ranging from 40 to 120 beats per minute. At the circled notes in the figure where the true tempo is slower than 67, the true tempo is equivalently translated to the doubled tempo with halved note lengths to find best matched model within prepared tempos between 67 and 120 beats/min.

E. Measure estimation

Estimation of measure and location of bar lines are also possible by using the HMM in a similar way. As depicted in Fig. 5,

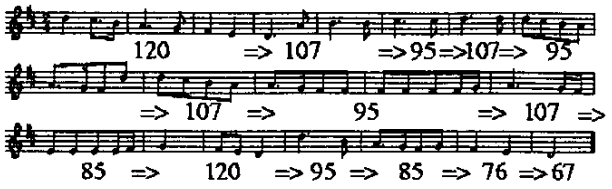

Fig. 7. Fluctuating tempo estimated by HMM. (G. F. Händel: "Joy to the world".)

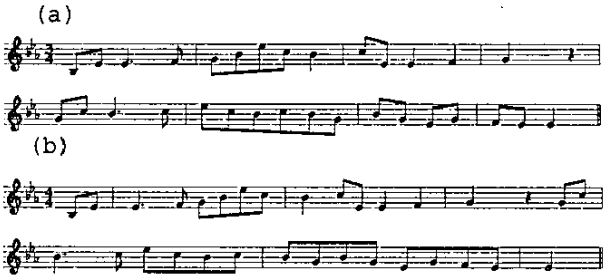

Fig. 8. An example of misrecognized measure. A: True score, B: Misrecognized but rather reasonable result.

one of multiple models representing different measures (e.g., 3/4 and 4/4) is found to yield higher likelihood for the given rhythm pattern. Each of these measure models has been trained with music data of the same measure.

Bar location is also estimated simultaneously in the same framework. As shown in Fig. 6, a special rhythm vocabulary model containing the starting rhythms and up-beat patterns in the first bar precedes the general rhythm vocabulary model consisting of general 2-beat-long rhythm patterns and bar lines used as an eternal loop. The Viterbi algorithm finds the optimal rhythm estimation with optimal bar locations.

In experimental evaluation of these models, 10 out of 10 testing music tunes of 4/4 measure and 8 out of 10 testing tunes of 3/4 measure were correctly recognized. Fig. 8(b) shows one of 2 misrecognized results which looks rather reasonable in the rhythm pattern sense. Correct recognition of this example requires higher knowledge such as: 3-bar phrase is rare in simple tunes.

IV. RHYTHM VECTOR APPROACH

A. Rhythm vector

We have discussed absolute note duration x_t as the observed feature in the HMM-based modeling. The use of relative durations of consecutive notes is discussed in this section.

Rhythm is primarily perceived as the relative length of consecutive notes. To define a tempo-free feature y_t instead of x_t, 3 consecutive notes durations are coupled to form a 3-dimensional vector $(x_{t-1}, x_t, x_{t+1})^T$ and normalized so that the sum of components is unity. By normalization, this tempo-free 3-dimensional "rhythm vector" $y_t = (y_{t-1}, y_t, y_{t+1})^T$ is mapped inside a triangular domain on a 2-dimensional plane.

Rhythm vector is considered to preserve tempo-free rhythmic intension in performance. Fig. 9(a) shows the plots of rhythm vectors calculated from the score shown as Fig. 9(c) and compared with Fig. 9(b) observed in human performances of the same score.

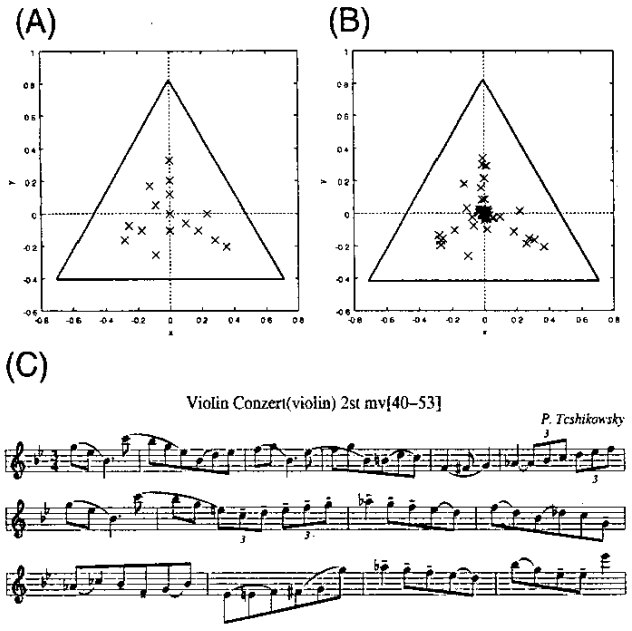

Fig. 9. Rhythm vectors plotted on a two-dimensional plane. A: Thoretical plots, B: Performance by human, C: Score.

B. HMM-based rhythm recognition

Replacing the observed absolute duration (IOI) of each note by "rhythm vector" in the HMM-based framework, tempo-free automatic transcription of MIDI data can be realized without preparing multiple HMMs for covering various tempos. A related idea was suggested in music pattern recognition purpose using the ratio of consecutive note durations to define a tempo-invariant encoding[7]. Hidden states correspond to the distinct points in Fig. 9(a) and integers are assigned to them to represent the state number. The observed vector distributing around the ideal position is modeled by a 3-dimensional pdf, $b_i(y_t)$, i denoting the state number. The transition probability, a_{ij}, from state i to state j is the probability of occurring of 4 consecutive notes, $x_{t-2}, x_{t-1}, x_t, x_{t+1}$, 2 of which in the middle are shared by the both states. These model parameters can be trained through the similar procedure as already stated for the 1-dimenstional observation case.

This method can be further improved by incorporating the absolute note durations. Since the rhythm vector is free from the absolute length of notes, the tempo is not uniquely determined. For example, when a rhythm vector is recognized as $1:2:2$, there are multiple possible note descriptions such as "Q H H" and "E Q Q" (H=half note, Q=quarter note, E=eighth note). Another problem is that one misrecognized rhythm vector may halve or double all following note lengths in decoding from the recognized rhythm vector sequence to the note description. These problem is avoided by giving prior information of intended approximate tempo or by including absolute note length in the feature vector for the HMM (i.e. hidden states corresponds notes for description). An alterntive solution to these problems is to select a path among N-best HMM traceback hypotheses with near-constant tempo. This can be easily realized by calculating the instantaneous tempo by the ratio of the observed IOI and the decoded note.

V. Discussion

Multi-voice music transcription: Though this paper has focused on transcription of single-voice music performed with a MIDI instrument, multi-voice music can be handled in the same framework. A chord can be identified as multiple notes started at the almost same timing (within a short time span) and overlapped in durations. As for multi-voice music such as counterpoint (fugues, cannons, etc.) can be also modeled by replacing the IOI along one voice by inter-onset interval between all voices. Such kind of "inter-voice rhythm" vocabulary can be acquired from a large amount of music data for training. After obtaining a single-voice transcription, it can be converted into a multi-voice music score taking into account the observed duration of each note.

Styles and genres: Obviously, the present approach relies on statistical characteristics of music both in rhythm vocabulary and n-gram approaches. This means music styles, genres, and composers can be reflected in these stochastic models to obtain better recognition abilities.

Weight adjustment: It should be noted that note duration modeling and rhythm vocabulary or n-gram modeling can be weighted depending on the purpose. If it is known beforehand that the player is not skillful in keeping the tempo and plays a relatively simple music, we can emphasize the rhythm vocabulary or n-gram constraints by giving a larger weight to a_{ij} in logarithmic likelihood calculation.

VI. Conclusion

We have discussed automatic rhythm recognition of MIDI signals of performed music through stochastic modeling note durations using HMM, the main technique for modern speech recognition. This can successfully estimate the sequence of intended note values (lengths), tempo (whether fixed and unknown or fluctuating), the time measure, and the bar locations all in the same modeling framework. Rhythm vector has been also introduced to enable tempo-free music transcription. Future works will include overall multi-stage integration of multi-voice transcription from MIDI signals covering tempo, time measure and bar location estimation, and integration with a multi-pitch detection technique for music transcription from the sound.

References

[1] R. Curtis: The Computer Music Tutorial, MIT Press, Cambridge, 1996.
[2] P. Desain and H. Honing: "The Quantization of Musical Time: a Connectionist Approach," Comp. Mus. J., Vol. 13, No. 3, pp. 56–66, 1989.
[3] N. Saito, M. Nakai, H. Shimodaira, and S. Sagayama, "Hidden Markov Model for Restoration of Musical Note Sequence from the Performance," Tech. Rep. Info. Proc. Soc. Jpn.
[4] T. Otsuki, N. Saito, M. Nakai, H. Shimodaira, and S. Sagayama, "Musical Rhythm Recognition Using Hidden Markov Model," Trans. Info. Proc. Soc. Jpn, Vol. 43, No. 2, pp. 245–255, 2002. (in Japanese)
[5] L. Rabiner, and B.-H. Juang: Fundamentals of Speech Recognition, Prentice-Hall, 1993.
[6] M. Hamanaka, M. Goto, H. Asoh, and N. Otsu, "A Learning-Based Quantization: Estimation of Onset Times in a Musical Score," Proc. SCI 2001, Vol. X, pp. 374–379, 2001. Vol. MUS-99, No. 106, pp. 27–32, 1999. (in Japanese)
[7] E.J. Coyle, I. Shmulevich, "A System Machine Recognition of Music Patterns," Proc. of ICASSP-98, pp. 3597-3600, 1998.