Proceedings of the International Symposium on Musical Acoustics, March 31st to April 3rd 2004 (ISMA2004), Nara, Japan

Maximum Likelihood Method for Estimating Rhythm and Tempo

Haruto TAKEDA[1], *Takuya NISHIMOTO* [1], *Shigeki SAGAYAMA* [1]

[1]Graduate School of Information Science and Technology, the University of Tokyo, Japan
{takeda,nishi,sagayama}@hil.t.u-tokyo.ac.jp

Abstract

This paper presents a rhythm recognition technique based on a probabilistic approach by utilizing generative model for timing information in expressive music performance. The problem of rhythm recognition including rhythm parsing and tempo tracking, is to retrieve information of rhythm and tempo from a sequence of observed note durations. Since performed note length deviates in real performance and decomposition of the duration into rhythm and tempo is not unique in general, this problem must be solved in a probabilistic approach. We formulate rhythm recognition as maximum a posteriori (MAP) state sequence estimation among a finite state network of Hidden Markov Models (HMMs). The structure of the proposed stochastic model is almost equivalent to a network model of HMMs used in continuous speech recognition technique. The most likely rhythm and tempo in this probabilistic model are obtained by use of an effective search algorithm, level building. Experimental evaluation using MIDI recordings of a classical music piece is also reported.

1. Introduction

Robust rhythm recognition techniques for expressive music performances are required because of its wide applicability including automatic transcription for music composition and arrangement, automatic content description of music material in large database, music information retrieval, performance analysis, score alignment and many multimedia applications. To perceive rhythm, average human usually uses many information including significant instrumental sound like drum or bass, code changes, knowledge of music genre; however, the technique discussed in this paper requires only temporal information. Therefore, the problem discussed here is how to retrieve intended rhythm and tempo from fluctuating note durations in expressively performed music.

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 with changing tempo. Transcribed scores are often far from the intended score.

On the other hand, it is easy for average listener to tap their hands in time with music, and trained musicians can easily transcribe performed (relatively simple) music even when the player changes the tempo intentionally for natural expression. The problem or rhythm recognition, often referred as “tempo tracking” or “rhythm parsing”, is thus considered essentially to involve rhythm pattern recognition utilizing *a priori* knowledge.

From this point of view, we previously introduced stochastic modeling based on Hidden Markov Model (HMM) for recognition of the rhythm pattern in given performed music, since rhythm recognition problem is analogous to continuous speech recognition[3]. In this framework, the sequence of note values, tempo (whether fixed and unknown, or fluctuating), the time sign, and the bar line allocation are successfully estimated from monophonic MIDI performances[2]. In this paper, the previous framework is extended to treat polyphonic music and combined with our other works[5], and unified viewpoint for rhythm recognition is given by MAP (Maximum A Posteriori) estimation.

Several research efforts exists to cope with the similar problem[6, 7, 8]. Our work differs from these in that only 2 consecutive onsets are treated in their probabilistic models while we use rhythm pattern in one measure as a unit of a model.

2. Stochastic Modeling

2.1. 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 a result, he can recognize the intended rhythm pattern. The inference is based on his knowledge about possible rhythm patterns acquired through musical 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 that generally underlies in music depending on genres, styles, and composers. To characterize possible rhythms, we use generative models for rhythm patterns. The “rhythm vocabulary” consists of all known rhythm patterns for a unit time (typ., one

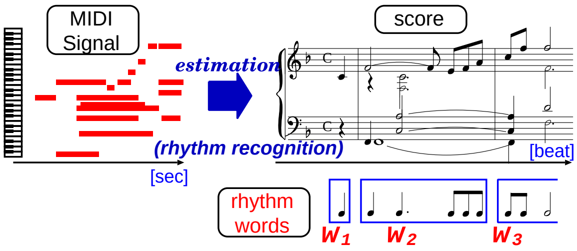

Figure 1: Rhythm parsing task is to find the best rhythm words sequence from timing information in expressive music performance

measure), that is refereed as "rhythm word". Rhythm recognition is the task of parsing performance timing information to rhythm words as shown in Figure 1. This model represents known rhythm patterns well whereby unknown patterns are substituted by similar existing patterns.

2.2. Note Lengths and Note Values

The observed duration (IOI, Inter-Onset Interval) x [sec] of note in the performance is related both to intended note value[1] q [beats] (in rhythm words) that appears in the score and to tempo variable $\tau(t)$ [sec / beat]. This is formulated by

$$x[\text{sec}] = \hat{\tau}(t)[\text{sec/beat}] \times q[\text{beats}] \qquad (1)$$

where average tempo is $\hat{\tau}(t) = \frac{\int_D \tau(t)dt}{|D|}$ [sec/beat] (D denotes the duration for the x, and $t \in D$).

2.3. Rhythm Recognition: MAP estimation

Rhythm recognition can be defined as a decomposition of IOIs $X{=}\{x_t\}_{t=1}^{N}{=}\{x_1, \cdots, x_N\}$ into tempo $T{=}\tau(t)$ and rhythm $Q{=}\ \{q_t\}_{t=1}^{N}{=}\{q_1,\cdots,q_N\}$. This is a kind of ill-posed problem since $\hat{Q}$ and $\hat{T}$ are not determined uniquely. In principle, any rhythm can be expressed in various ways, e.g. twice note values and half tempo gives the same note duration in Eq. 1. Furthermore, fluctuation of tempo and rhythm cannot be completely separated. Decomposition is possible only in a probabilistic sense assuming that T is constant or slowly changing (at least within phrases), and that Q often fit to common rhythm patterns expressed by a sequence of rhythm words $W{=}\{w_m\}_{n=1}^{M}{=}\{w_1, \cdots, w_M\}$ in rhythm vocabulary. Therefore, the approach based on MAP (Maximum A Posteriori) must be applied to find "most likely rhythm patterns" $\hat{W} = \{\hat{w}_m\}_{m=1}^{M}$ and "most likely intended tempo" $\hat{T} = \hat{\tau}(t)$, given a sequence of observed note lengths series, X. This is formulated as follows:

$$\{\hat{W}, \hat{T}\} = \underset{\{w_m\}_{m=1}^{M}, T}{\text{argmax}} \ P(W, T|X)$$

Length of rhythm words, M, is also variable in the search.

Since maximizing $P(W,T|X)$ for given X is equivalent to maximizing $P(X|W,T)P(W,T)$ according to the Bayes theorem, finding most likely $\hat{W}$ and $\hat{T}$ among all possible W is formulated as,

$$\{\hat{W}, \hat{T}\} = \underset{W,T}{\text{argmax}}\, P(X|W,T)P(W,T). \qquad (2)$$

Our goal is to separate rhythm and tempo by iterating estimation of the two. In the rest of this paper, we describe a method of rhythm estimation to give a initial value for the first step of this iterations. Assuming that rhythm W and tempo T is independent, $P(W,T)$ is approximated by $P(W)P(T)$. Then, Eq. 2 can be written as

$$\hat{W} = \underset{W}{\text{argmax}}\, P(X|W)P(W)P(T) \qquad (3)$$

3. HMMs for Polyphonic Music

3.1. HMMs for Rhythm Words

Suppose that consecutive n IOIs $x_t, \cdots, x_{t+n-1}$ and a rhythm word $w_i = \{q_1, \cdots, q_{K(i)}\}$ are given, where $k(i)$ denotes the number of notes included in the rhythm word w_i. When several notes are intended to play simultaneously in a polyphonic music, short time IOIs (ideally 0) are observed like x_1 in Figure 2. These IOIs correspond to the same note value q in a rhythm word w. We model this situation by using HMM (hidden Markov Model) and associate note value and observed IOIs. As shown in Figure 2, HMM states correspond to note values in a rhythm word, and IOIs are output value from state transition.

In this HMMs, probabilities are given for each state transition and transition output. Auto-transition probability are given by

$$b_{ss}(x) \propto \frac{1}{1 + e^{\beta(x-\mu)}} \qquad (4)$$

We set $\mu = 50ms$ as around 50ms difference in onset time are common in piano and ensemble performance according to the studies of chord asynchrony. $a_{s(t)s(t+1)}$ denotes a probability to transit from state $s(t)$ to state $s(t+1)$.

3.2. Probability for Tempo Variations

Fluctuation of tempo is also treated with probabilities. Since tempo varies slowly in the most case, and statistics of variation of tempo is expected to distributed around 0. Average tempo $\bar{\tau}$ of rhythm word w_i is

$$\bar{\tau}_i = \frac{x_t + \cdots + x_{t+n-1}}{q_1 + \cdots + q_{K(i)}} \qquad (5)$$

[1] Note values are nominal information of notes. For example, if a note value of quarter note is set 1, that of half note is 2 and that of eighth note is $1/2$.

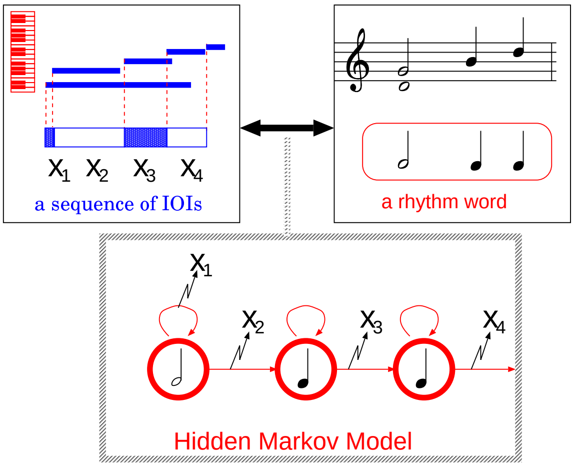

Figure 2: Observed IOIs and rhythm words are associated in the framework of Hidden Markov Models (HMMs)

and a sequence of average tempo $\{\bar{\tau}_m\}_{m=1}^{M}$ is modeled by a probabilistic distribution for their differences by

$$\bar{\tau}_{n+1} - \bar{\tau}_n \sim N(0, \sigma) \tag{6}$$

where $\bar{\tau}_m$ denotes average tempo of m-the rhythm word w_m in a sequence. Then, $P(T)$ is obtained from a sequence of the average tempo $\{\bar{\tau}_m\}_{m=1}^{M}$ in a rhythm words sequence $\{w_m\}_{m=1}^{M}$.

3.3. Rhythm Vectors: Tempo-invariant Features

From the assumption that tempo is almost constant in a rhythm word, the proportion of note lengths x in a rhythm word w_i is almost independent of tempo $\tau(t)$ according to Eq. 1. Therefore, we introduce K(i) dimensional *rhythm vector* $\boldsymbol{r}_t$=$(r_t)_{j=1}^{K(i)}$, whose component represents proportion of IOIs as follows:

$$r_t^j = \frac{x_{t+j}}{x_t + \cdots + x_{t+n-1}} \tag{7}$$

where x_j corresponds to state transition output of a HMM [2] By this definition, a rhythm vector sequence R=$\{\boldsymbol{r}_n\}_{n=1}^{M}s$ is computed from an an observed IOI sequence X=$\{x_t\}_{t=1}^{N}$.

We assume that fluctuation of rhythm vectors is observed with different probability and this stochastic variable $\boldsymbol{r}$ follows the normal distribution $N(\boldsymbol{\mu}, \Sigma)$. Therefore, probability that a rhythm word w_i is performed as a sequence of IOIs $\{x'_t\}_{t'=t}^{t+n-1}$ is given by

$$\begin{aligned} & P(x_t, \cdots, x_{t+n-1}|w_i) \\ & = \prod_{t'=t}^{t+n-1} a_{s(t')s(t'+1)} b_{s(t')s(t'+1)}(x'_t) c(\boldsymbol{r}_{t'}|w_i) \end{aligned} \tag{8}$$

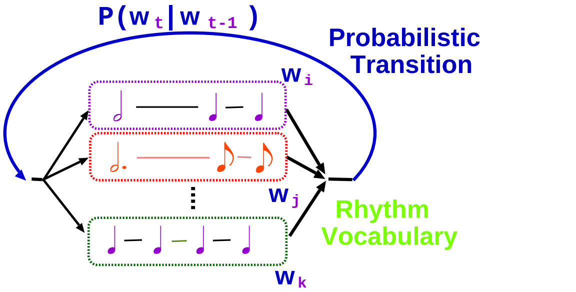

Figure 3: Probabilistic grammar is introduced to rhythm vocabulary network

where $c(\boldsymbol{r})$ is the probability density function of rhythm vector.

In our model, consecutive IOIs that corresponds to a rhythm word can be divided into two factors, i.e., tempo $\bar{\tau}$ and rhythm vector $\boldsymbol{r}$. By use of probabilistic model above mentioned, probability that a sequence of all IOIs $X = \{x_t\}_{t=1}^{N}$ is observed for a given word rhythm sequence $W = \{w_m\}_{m=1}^{M}$, $P(X|W)P(T)$ is obtained by product of Eq. 6 and Eq. 8 through the corresponding HMM network.

3.4. Probabilistic Grammar for Rhythm Words

Similar to the n-gram language model often used in speech recognition, the probability of rhythm word sequence $W = \{w_m\}_{m=1}^{M}$ can be approximated by

$$P(W) \approx P(w_1, \cdots, w_{n-1}) \prod_{t=n}^{M} P(w_t|w_{t-1}, \cdots, w_{t-n+1}) \tag{9}$$

For example, HMM network of rhythm words bigram model (n=2) is shown in Figure 3. Conditional probabilities can be obtained through statistical training using already composed music scores.

3.5. Search in a network of HMMs

To obtain the most likely rhythm pattern in Eq. 3, likelihoods of each rhythm words sequence hypothesis must be calculated. Then, the problem is a search process in a network of HMMs that consists of state transition network. As shown in Figure 4, optimal state transition can be obtained by Viterbi search algorithm and the optimal HMMs sequence can be obtained by level building.

3.6. Experimental Evaluation

The proposed method was evaluated by using performance data of J. S. Bach's "Fuga in C minor"[3] recorded in the MIDI format, which were played by 2 piano players 2 times. 14 kinds of note values (whole note, quar-

[2] In Figure 2, elements of the rhythm vector is obtained from x_2,x_3,x_4.

[3] BWV847, Well-Tempered Clavier Book I

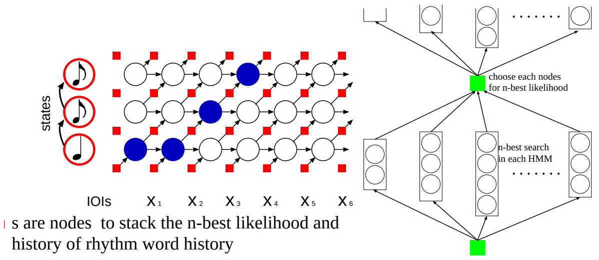

Figure 4: An optimal sequence of rhythm word was obtained by network search of rhythm words, in which the optimal state sequences are also found by Viterbi search algorithm

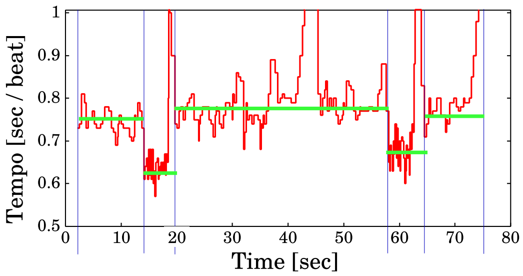

Figure 5: In expressive music performance, tempo fluctuates, and sometimes it changes suddenly.

Table 1: Accuracy of Rhythm words for MIDI performance data of J. S. Bach's "Fuga in C minor"

training data	rhythm word	word accuracy
Fuga	7	100.0%
6 pieces	52	96.7%

ter note, etc.) were treated in the proposed HMMs and rhythm word bigram model. We constructed 2 bigram models; one is constructed by 7 rhythm words obtained from statistics of "Fuga" and another one consists of 52 rhythm words through training from 6 classical pieces. We set $\beta = 80$ in Eq. 4. Accuracy of rhythm words was calculated by

$$\text{Acu.} = \frac{M - D - I - S}{M}$$

where I, S, D denotes error of insertion, substitution, and deletion for each. M is the number of rhythm words in the original score. Accuracy of rhythm words is listed in Table 1. An errors of substitution were observed by taking IOIs for sixteenth notes as simultaneous onsets.

3.7. Tempo Fitting

Since rhythm recognition problem is given in Eq. 2, tempo should be estimated as well as rhythm. Tempo $\tau(t)$ in Eq. 1 is obtained from estimated q. We have already proposed a method by fitting this observed tempo $tau(t)$ with a tempo model $\tau(t)$ to estimate time points of tempo change as shown in Figure 5 which Schumann's"Fürchtenmachen"[4][9]. The method of detecting tempo change points and the proposed model in this paper can be integrated by taking the observation unit of tempo in each rhythm word.

4. Conclusion

We have discussed rhythm recognition of MIDI signals of performed music through stochastic modeling of note durations using HMMs, the key technique for modern speech recognition. This can successfully estimate the sequence of intended rhythm from polyphonic music performances.

Future works include introducing global modeling of tempo and probabilistic model of (harmony) information, and integration with a multi-pitch detection technique for music transcription from the audio signals.

References

[1] R. Curtis. "The Computer Music Tutorial, MIT Press," Cambridge, 1996.

[2] N. Saitou, M. Nakai, H. Shimodaira, S. Sagayama, "Hidden Markov Model for Restoration of Musical Note Sequence from the Performance," Proceedings of the Joint Conference of Hokuriku Chapters of Institutes of Electrical Engineers, Japan, 1999, F-62, p.362, Oct 1999 (in Japanese).

[3] L. R. Rabiner, B. H. Juang, "An Introduction to Hidden Markov Models," ASSP magazine, pp. 4–16, 1986.

[4] H. Takeda, T. Otsuki, N. Saito, M. Nakai, H. Shimodaira, S. Sagayama, "Hidden Markov Model for Automatic Transcription of MIDI Signals," Proc. 2002 IEEE Workshop on Multimedia Signal Processing (MMSP), 2002.

[5] H. Takeda, T. Nishimoto, S. Sagayama, "Automatic Rhythm Transcription for Multiphonic MIDI Signals," Proceeding of the 4th International Symposium of Music Information Retrieval (ISMIR), pp. 263–264, 2003.

[6] A. Cemgil, B. Kappen, P. Desain, H. Honing, "On tempo tracking: Tempogram Representation and Kalman filtering," Journal of New Music Research, 2000.

[7] M. Hamanaka, M. Goto, H. Asoh, and N. Otsu, " Learning-Based Quantization: Estimation of Onset Times in a Musical Score," Proceedings of the 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001), Vol. X, pp. 374–379, 2001.

[8] C. Raphael, "Automated Rhythm Transcription," Proceeding of the 2nd International Symposium of Music Information Retrieval (ISMIR), pp. 99–107, 2001.

[9] H. Takeda, T. Nishimoto, S. Sagayama, "Estimation of Tempo Variations in Performed MIDI Data Signals using Rhythm Vectors," Information Processing Society of Japan (IPSJ) SIG Technical reports, 2003-MUS-51, pp.59–64, 2003. (in Japanese)

[4] a piano piece from "Kinderzenen", Op. 15, No.11