

RHYTHM AND TEMPO RECOGNITION OF MUSIC PERFORMANCE FROM A PROBABILISTIC APPROACH

Haruto Takeda Takuya Nishimoto Shigeki Sagayama
Graduate School of Information Science and Technology
The University of Tokyo

ABSTRACT

This paper concerns both rhythm recognition and tempo analysis of expressive music performance based on a probabilistic approach. In rhythm recognition, the modern continuous speech recognition technique is applied to find the most likely intended note sequence from the given sequence of fluctuating note durations in the performance. Combining stochastic models of note durations deviating from the nominal lengths and a probabilistic grammar representing possible sequences of notes, the problem is formulated as a maximum *a posteriori* estimation that can be implemented using efficient search based on the Viterbi algorithm. With this, significant improvements compared with conventional “quantization” techniques were found. Tempo analysis is performed by fitting the observed tempo with parametric tempo curves in order to extract tempo dynamics and characteristics of performance to use. Tempo-change timings and parameter values in tempo curve models are estimated through the segmental *k*-means algorithm. Experimental results of rhythm recognition and tempo analysis applied to classical and popular music performances are also demonstrated.

keywords: rhythm recognition, hidden Markov models, tempo analysis, segmental *k*-means algorithm, continuous speech recognition framework, *n*-gram grammar

1. INTRODUCTION

Techniques for restoring music score information from musical performances are useful in content-based music information retrieval (MIR). This paper concerns a method for estimating the temporal factors of a score from given musical performance data using rhythm recognition and tempo analysis.

Music score information plays an important role in MIR because of its flexibility and its compactness compared to audio signals. A fast query search by melody or rhythm pattern is possible using the score data stored in a database. In addition, score data is flexible against

musical conversion like transposition (key changes). Utilizing these features, similarity, for instance, can be efficiently calculated between the search query and music contents. Large music databases of audio contents, however, are typically not associated with score information corresponding to the contents. Thus needs for technique to obtain or restore score information from the audio signals. The technique can also be applied to feature extraction for tagging the meta data in MPEG4 contents.

Currently, most methods for restoring sheetmusic score from music audio signals consists of two processing stages. First, spectrum analysis of audio signals is done to detect pitch frequency and onset timing of each note event in the audio signals. The result can be shown in a piano-roll display and can usually be recorded in the standard MIDI (Musical Instrument Digital Interface) file (SMF). In the next step, score information notated by symbols, is restored from the SMF data obtained from the first processing stage. Though the audio signal analysis process is not a trivial problem, excellent performance is attained by several recent efforts, such as “specmurt analysis” [1] which converts spectrogram into a piano-roll-like display nearly equivalent to MIDI data. Alternatively, music can be played with MIDI instruments such as electronic piano that directly produces MIDI signals, the audio signal processing step can be skipped.

Now, the paper will focus on the latter process, assuming that the music performance data is given as a MIDI signal. The methods described in this paper can be applied to any performance data which contain note onset timing information.

Quantization, the conventional method for rhythm extraction from MIDI performance, does not work well for expressive music as shown in Fig. 2. Since human performers changes tempo and note lengths both intentionally and unintentionally to make their performances more expressive, the note lengths deviates so much from the nominal note lengths intended by the performer, that simple quantization of note lengths can not restore the intended note length and often results in an undesired (funny) score.

On the other hand, when human listen to the music, they can usually perceive its rhythmic structure and clap their hands to the beat of the music. If they have acquired musical knowledge through their musical training, they can even give a reasonable interpretation of the rhythm as

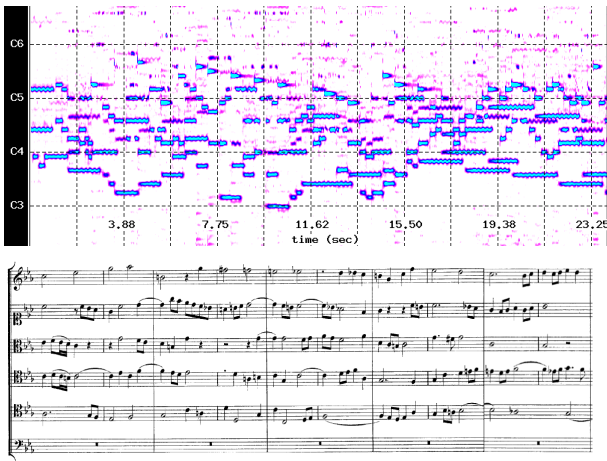


Figure 1. A piano-roll-like result of “specmurt analysis” (top) applied to a real music signal of “J. S. Bach: Ricer-care à 6 aus das Musikalische Opfer, BWV 1079” (score at bottom) performed by a flute and strings, excerpted from the RWC music database [11].

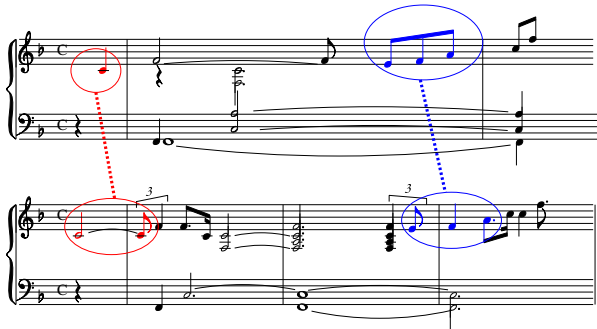


Figure 2. The result of quantization of a MIDI signal by commercial software (lower) compared to the original score (upper) of “Träumerei” played on an electronic piano.

a note sequence since they know which rhythm patterns are more likely to appear among all possible rhythms. They do not quantize note lengths they hear, but instead, recognize a sequence of the performed note lengths as a rhythm pattern. In summary, the rhythm is something not to quantize but to recognize. Therefore, to estimate rhythm patterns from performed note lengths, we focus on an algorithm to recognize the rhythm patterns from the view point of speech recognition.

We proposed a new rhythm recognition approach[3, 4] in 1999 utilizing probabilistic modeling which is often employed in modern continuous speech recognition (CSR) technology from our viewpoint of strong analogy between rhythm recognition and speech recognition. Speech recognition[2] takes a feature vector sequence as input and outputs the recognized word sequence, while rhythm recognition takes the note length sequence as input and outputs the rhythm patterns. In the proposed model, both appearance of rhythm patterns and deviation of note length are associated with probability to evaluate how likely hypothesized rhythm patterns are really in-

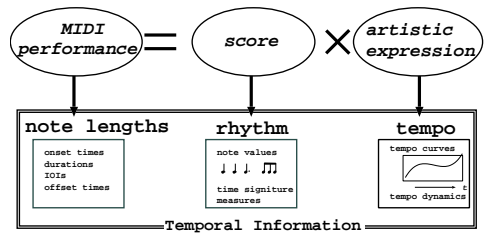


Figure 3. Temporal information of performance data consists of score information (rhythm) and artistic expression (tempo).

tended in the given performance. In this approach, we defined a probabilistic vocabulary of rhythm words trained with a music database. The rhythm recognition problem was formulated as a connected rhythm word recognition and solved by a continuous speech recognition algorithm. This framework simultaneously enabled bar line allocation by adding “up-beat rhythm” words, beat recognition by preparing two-beat vocabulary and three-beat vocabulary connected in parallel, and tempo estimation both for changing tempo and unknown tempo. In this approach, the model parameter values can be optimized through stochastic training, and rhythm recognition can be performed with an efficient search algorithm.

There have also been several efforts for rhythm recognition based on probabilistic modeling [5, 6] to estimate note values or beats although the time signature has to be given before recognition, and *a priori* probabilities of rhythm pattern is not taken into account. We discuss our approach to rhythm recognition in Section 2.

In addition to rhythm, tempo is another important factor for MIR. Though they are both related to temporal factors in music, rhythm is primarily related to microscopic changes in consecutive note lengths and tempo is more related to macroscopic and slow changes. As shown in Fig. 3, these two factors are coupled to yield each of observed note durations. Tempo sometimes changes rapidly like *Adagio* to *Allegro*. The local tempo fluctuations within phrase depend on music genre, style and performers. Tempo often expresses artistic characteristics of the performance, while rhythm expresses the intended score. If these factors are separately extracted from music performance, they may be effective for content-based music search like “music that has overture and allegro”, or “performance playing that phrase very slowly”.

There are some researches that dealt with performed tempo for analyzing the performance characteristics. While previous works of tempo analysis includes visualization of performance [7] and comparison of performers background (jazz and classical) by periodic statistics of tempo [8], our objective is to extract information that characterize the performances including tempo changes and tempo handling in phrases. We propose a tempo analysis method by estimating partly smooth and continuous “tempo curves.” It will be discussed in Section 3.

2. RHYTHM RECOGNITION

2.1. Rhythm Vocabulary

Extending the analogy between rhythm recognition and speech recognition, we introduce a “rhythm vocabulary” in order to construct a probabilistic model for rhythm recognition. Comparing human knowledge about rhythm patterns to a stochastic language model in modern CSR technology, rhythm patterns can be modeled as a stochastic note generating process. This model generates the note sequence of a rhythm pattern associated with a probability that varies on music genres, styles, and composers of a “rhythm vocabulary”. The “rhythm vocabulary” consists of units (this time, one measure) called “rhythm word”. A rhythm vocabulary and a grammar of rhythm words can be obtained through stochastic training using existing music scores.

One advantage of using rhythm words for modeling rhythm patterns is that meter information can be estimated simultaneously along with notes. Thus, the locations of bar lines in a score correspond with the position of boundaries in a rhythm word sequence. Time signature is also determined by investigating sum of note values in estimated rhythm words.

2.2. Probabilistic Grammar for Rhythm Vocabulary

Similar to language model of CSR, n -gram model of rhythm words is used for a grammar of rhythm vocabulary. That is, the probability of a rhythm word sequence $W = \{w_m\}_{m=1}^M$ is approximated by cutting out the history of rhythm word appearance,

$$P(W) = P(w_1, \dots, w_{n-1}) \cdot \prod_{m=n}^M P(w_m | w_{m-1}, \dots, w_{m-n+1}) \quad (1)$$

Conditional probabilities can be obtained through statistical training using previously composed music scores.

The n -gram model reflects the local features of the music passage, but does not the global structure including repetition of rhythm patterns. As is often the case with CSR, unknown rhythm patterns in the vocabulary is substituted with similar existing patterns. To obtain more reliable values for model parameters, linear interpolation or other techniques commonly used for language model can be applied.





2.3. Nominal Relation of Temporal Information

The observed duration (IOI, Inter-Onset Interval) x [sec] of note in the performance is related both to the note value¹ (time values) q [beats] in score and the tempo τ [BPM] (beats per minute) as follows:

$$x[\text{sec}] = \frac{60[\text{sec}/\text{min}]}{\tau[\text{beats}/\text{min}]} \times q[\text{beats}] \quad (2)$$

¹ “Note values” are nominal length of notes. For example, if a note value of quarter note is defined as 1[beat], that of half note is 2[beats] and that of eighth note is 1/2[beat].

Table 1. Rhythm word examples and their probabilities obtained thorough stochastic training.

rhythm words w	$P(w)$
	0.1725
	0.1056
	0.0805
	0.0690
.....

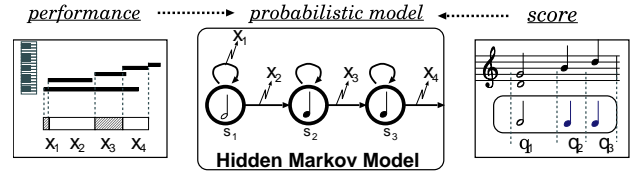


Figure 4. Observed IOIs and rhythm words are associated in the framework of Hidden Markov Models (HMMs).

2.4. Modeling Rhythm Words using HMMs

A rhythm word and a sequence of deviating IOIs are probabilistically related using a Hidden Markov Model (HMM) [9].

Suppose that consecutive n IOIs, x_k, \dots, x_{k+l} , and a rhythm word, $w_i = \{q_1, \dots, q_{S_i}\}$, are given, where S_i denotes the number of notes contained in the rhythm word w_i . When several notes are intended to be played simultaneously in polyphonic music, short time IOIs (ideally 0) are observed, such as x_1 in Fig. 6. These IOIs correspond to the same note value q in a rhythm word w_i . We model this situation by using the HMM and associate note values and observed IOIs. As shown in Fig. 6, HMM states correspond to note values in a rhythm word, and IOIs are output value from state transitions.

In the HMMs, probabilities are given to each state transition and transition output. Probability of observing x is modeled with a zero-mean normal distribution at auto-transition of state s . $a_{s(k)s(k+1)}$ denotes a probability to change from state $s(k)$ to state $s(k+1)$. Self-transition probability $a_{s,s}$ corresponds to the times of stay in state s , that is, the number of notes simultaneously played in the state, whose expectation is given by $\frac{1}{1-a_{s,s}}$. Values of $a_{s,s}$ are automatically determined with statistics of score, as shown in Fig. 5. Variation of IOIs that corresponds to note values q is assumed to distribute normally with means $\frac{60}{\tau} \cdot q_s$ [sec] and variance σ^2 , where τ is the average tempo of the previous rhythm word described in 2.5. This corresponds to the output probability of state transition $b_{s,s+1}(x)$.

Therefore, the probability that a rhythm word w_i is performed as a sequence of IOIs $\{x_{k'}\}_{k'=k}^{k+l}$ is given by

$$P(x_k, \dots, x_{k+l} | w_i) = \prod_{k'=k}^l a_{s(k')s(k'+1)} b_{s(k')s(k'+1)}(x_{k'}) \quad (3)$$

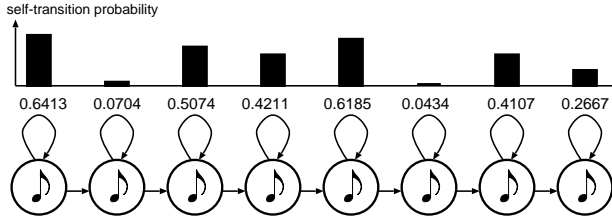


Figure 5. An example of stochastic training of state transition probabilities: States of strong beats have higher probability of self-transition than states of weak beats.

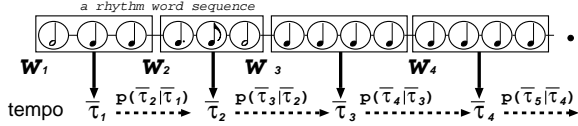


Figure 6. Tempo tracking in each rhythm word using probability of tempo variations.

2.5. Probability of Tempo Variations

The fluctuation of the performed tempo is also treated with probabilities. Since we do not have it a priori knowledge about the tempo variation specific to the given performance, we simply assume that a tempo of a measure is close to that of the previous measure. The average tempo $\bar{\tau}$ in a measure with rhythm word w_i is calculated using Eq. (2) by

$$\bar{\tau} = \sum_{k'=k}^l x_{k'} / \sum_{s=1}^{S_i} q_s$$

We give conditional probability for consecutive average tempo $P(\bar{\tau}_m | \bar{\tau}_{m-1})$ by assuming that the difference $\log \bar{\tau}_m - \log \bar{\tau}_{m-1}$ in log scale distributes normally with mean 0.

Then, the probability that an IOI sequence X is observed for a given word rhythm sequence W , $P(X|W)$ is obtained from the product of Eq. (3) and the probability of tempo variations

$$P(X|W) = \prod_{m=1}^M P(x_{l(m)}, \dots, x_{l(m+1)-1} | w_m) P(\bar{\tau}_{m+1} | \bar{\tau}_m) \quad (4)$$

where $x_{l(m)}$ denotes the first IOI in the m -th rhythm word.

2.6. MAP Estimation for Rhythm Recognition

By integrating these probabilistic models, rhythm recognition can be formulated as a MAP estimation problem. Using a rhythm vocabulary, rhythm recognition can be defined to find the “most likely rhythm patterns” \hat{W} for a given IOI sequence X . According to the Bayes theorem,

$$\hat{W} = \operatorname{argmax}_{\{w_m\}_{m=1}^M} P(W|X) = \operatorname{argmax}_{\{w_m\}_{m=1}^M} P(X|W)P(W) \quad (5)$$

where the number of rhythm words, M , is also variable in the search. In our model, Eqs. (1) and (4) are substituted with Eq. eq:argmax W.

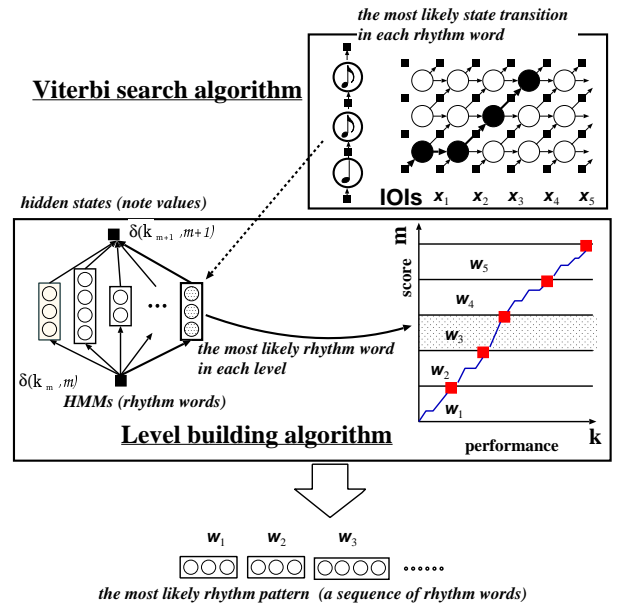


Figure 7. Network search to find the optimal rhythm-word sequence and the optimal state sequence using the Viterbi search algorithm.

2.7. Search Algorithm

Finding the most likely rhythm word sequence in Eq. (5) is a search process in a network of HMMs that, in turn, each consist of state transition networks. Several search algorithm developed for CSR can be applied for this purpose, since models of both recognitions share the common hierarchal network structure.

This time, we implemented the search using the Level Building algorithm [10]. In the following algorithm, $\delta(t, m)$ stands for the highest cumulative likelihood for the t th IOIs with m rhythm words. The Viterbi algorithm is used for calculating $\delta(t, m|w)$.

```

for m=1 to maximum_number_of_bar_lines
  for every w' in rhythm_vocabulary
    for t=1 to num_of_notes
       $\delta(t, m|w_1, \dots, w_{m-1}, w')$ 
        =  $\max_{t'} \delta(t', m) + d(t', t|w')$ 
    for every t=1 to T
       $\delta(t, m) = \max_w \delta(t, m|w)$ 
   $\hat{W} = \operatorname{argmax}_m \delta(T, m)$ 

```

2.8. Experimental Evaluation

The proposed method was evaluated with performance data played by human with electronic piano² and recorded in SMF as listed in Table 2. Data M1 consists of relatively simple rhythm patterns and was played with nearly

² YAMAHA Clavinova.

Table 2. Test data for rhythm recognition experiments.

data ID	music piece
M1	J. S. Bach: Fuga in c-moll, BWV847. from Das wohltemperierte Klavier, Teil 1.
M2	R. Schumann: “Träumerei” from “Kinderszenen,” Op. 15, No. 7.
M3	L. v. Beethoven: 1st Movement of Piano Sonata, Op. 49-2.
M4	W. R. Wagner: “Brautchor” from “Lohengrin”
M5	The Beatles: “Yesterday”
M6	The Beatles: “Michelle”

Table 3. 3 conditions of constructing rhythm vocabulary.

condition	training data	#rhythm words
closed 1	each of testing data (M1~M6)	14,10,12,16,9,8
closed 2	22 pieces including testing data	162
open	16 pieces excluding testing data	139

constant tempo. On the other hand, the tempo of M2 (Träumerei) changed much in the performance according to the tempo indication of *rit* and the performers’ individual expression. M3 tends to be played with constant tempo, but rhythm patterns include eighth and triplet eighth notes.

To construct of rhythm vocabulary, a bigram model ($n = 2$ in Eq. (1)) was trained under 3 conditions listed in Table 3. The first condition “closed 1” is the most specific condition of the three, where the rhythm vocabulary has been extracted from the testing music material. The second condition “closed2” shares the same rhythm vocabulary extracted from all testing materials. Under 3rd condition “open”, the model has been acquired from 16 music pieces different from testing materials. In this case, some rhythm patterns in the testing music may be missing in the trained vocabulary.

Accuracy of note values q for each IOI x was evaluated by $\frac{N-S}{N}$, where N is the number of IOIs and S denotes the number of misrecognized IOIs. Also, accuracy both of rhythm-words in each measure and of locations of bar lines were evaluated by:

$$\text{Acc} = \frac{N - D - I - S}{N}$$

where I , S , D denote insertion, substitution and deletion errors, respectively, and N is the number of measures in the original score.

Tables 4, 5 and 6 show results of rhythm recognition significantly superior to the note value accuracy obtained by the quantization method: 14.4–18.8%. A typical misrecognition is due to failure to track tempo in several parts where the tempo changes much within a measure as a result of the indication of *rit*. or performer expression. Since we modeled tempo as constant within a rhythm word, the HMM could not adapt to such a rapid tempo change. Another typical misrecognition was that eighth notes were

Table 4. Accuracy of note value q of IOI x in the performance [%].

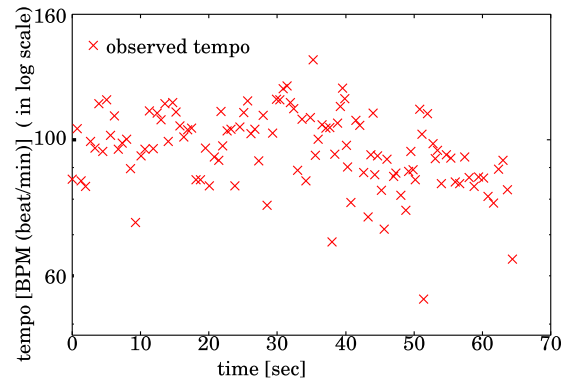
model	M1	M2	M3	M4	M5	M6	ave.
closed 1	99.8	99.7	100	100	100	100	99.9
closed 2	98.5	95.7	100	100	93.7	94.2	96.4
open	89.8	62.3	80.7	48.3	90.0	90.6	76.9

Table 5. Accuracy of rhythm word w in rhythm score [%].

model	M1	M2	M3	M4	M5	M6	ave.
closed 1	100	95.8	100	100	100	100	99.3
closed 2	93.3	88.0	100	100	70.8	96.5	91.4
open	60.0	46.0	68.4	18.8	45.8	86.2	54.1

Table 6. Accuracy of bar line allocations [%].

model	M1	M2	M3	M4	M5	M6	ave.
closed 1	100	99.7	100	100	100	100	99.9
closed 2	100	83.3	100	100	87.5	100	93.7
open	46.6	50.0	78.9	60.4	54.1	100	65.0

**Figure 8.** Tempo [BPM] for IOIs in piano performance of “Michelle” by The Beatles.

sometimes misrecognized as triplets. Recognition performance degraded for “open-data” training cases most possibly due to insufficient training data.

3. TEMPO ANALYSIS

3.1. Multilayer Tempo Characteristics

After rhythm recognition of the performed music data, instantaneous local tempo $\tau_k = \frac{x_k}{q_k}$ can be calculated from the observed IOI x and estimated note value q according to Eq. (2). As the estimated instantaneous local tempo, however, fluctuates almost randomly as shown in Fig. 8, tempo analysis is necessary to extract the “true” tempo underlying behind the observed tempo.

We assume that musical performances contain hierarchical (multilayer) tempo-related factors with different

time scales. For example, each measure contains rhythmic characteristics based on traditional music styles, such as Wiener Waltz, Polonaise, etc. The melody phrase may be characterized by the performers' articulations or tempo control styles according to their artistic expression. Music works are often composed of several parts, each with its own different tempo indication, and include drastic tempo changes in the music pieces.

Our strategy for obtaining tempo characteristics of each hierarchical structure is to fit the performed tempo within time segments to a tempo pattern by optimizing the model parameters, and also to cluster several consecutive measures in order to form tempo curves. In the proposed model, slow changes in tempo are modeled as a tempo curve in each segment, while drastic tempo changes are dealt as boundaries between different segments. The rhythm recognition discussed in Section 2 provides a method to estimate the note sequence given a sequence of IOIs in Eq. (2). In this section, we provide a method for tempo analysis by detecting timings of tempo changes and by fitting a tempo curve to partial music phrases.

3.2. Formulating the Tempo Curve

Since the sequence of local tempos $\{\tau_k\}_{k=1}^N$ includes fluctuations and deviations in the performance, we model the performed tempo with multiple concatenated smooth tempo curves where a tempo curve $\tau(t|\boldsymbol{\theta})$ is a continuous function of time t [sec] with parameters $\boldsymbol{\theta}$ and modeled by polynomial function in the logarithmic scale, i.e.,

$$\log \tau(t|\boldsymbol{\theta}) = a_0 + a_1 t + a_2 t^2 + \dots + a_P t^P \quad (6)$$

with parameters $\boldsymbol{\theta} = \{a_0, \dots, a_P\}$.

Now, we assume that the difference between the observed tempo τ_k and the modeled tempo $\tau(t_n|\boldsymbol{\theta})$ at the n -th onset time on the tempo curve, i.e., $\epsilon_k = \log(\tau(t_k|\boldsymbol{\theta})) - \log(\tau_k)$, can be regarded as a probabilistic deviation from a normal distribution with mean 0 and variance σ^2 . Therefore, the simultaneous probability of deviations of all notes is given by:

$$p(\epsilon_1, \dots, \epsilon_N) = \prod_{k=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log \tau_k - \log \tau(t_k|\boldsymbol{\theta}))^2}{2\sigma^2}\right) \quad (7)$$

3.3. Probability of Tempo Changes

In this paper, we assume that tempo is nearly constant within segments and sometimes changes drastically between them. We model the probability of changing tempo between consecutive segments by:

$$P(\bar{\tau}_k, \bar{\tau}_{k+1}) = 1 - \exp\left(-\frac{(\bar{\tau}_k - \bar{\tau}_{k+1})^2}{2\sigma^2}\right) \quad (8)$$

where $\bar{\tau}_k$ is the average tempo within a segment in the k -th tempo model,

$$\bar{\tau}_k = \frac{\sum_{k=S_r}^{S_{r+1}-1} \tau_k x_k}{\sum_{k=S_r}^{S_{r+1}-1} x_k}$$

and S_k indicates the index of the first note of the k -th segment. Eq. (8) yields a probability of 0 when tempo stays the same value $\bar{\tau}_k = \bar{\tau}_{k+1}$.

3.4. MAP Estimation of Tempo Analysis

We use the maximum *a posteriori* probability as a criterion for optimizing the model in order to find the best fitting tempo patterns and to detect the timings of tempo changes. In other words, given the sequence of onset timings of a performance and the corresponding note values, the most likely tempo curves are estimated. With the Bayes theorem, the tempo analysis can be written as:

$$\hat{T} = \underset{T}{\operatorname{argmax}} P(T|X, Q) = \underset{T}{\operatorname{argmax}} P(X, Q|T)P(T) \quad (9)$$

where T denotes the tempo curve, X the performance, and Q the score information. This time, $P(X, Q|T)$ is given in Eq. (7), and $P(T)$ in Eq. (8), and by taking logarithm of them, Eq. (9) is found equivalent and can be used in finding concatenated tempo curves $\tau(t|\hat{\boldsymbol{\theta}}_1, \dots, \hat{\boldsymbol{\theta}}_R, \hat{S}_1, \dots, \hat{S}_{R-1})$ with the parameters estimated by:

$$\begin{aligned} & \{\hat{\boldsymbol{\theta}}_1, \dots, \hat{\boldsymbol{\theta}}_R, \hat{S}_1, \dots, \hat{S}_{R-1}\} \\ &= \underset{\{\boldsymbol{\theta}\}_{k=1}^R, \{S_r\}_{r=1}^{R-1}}{\operatorname{argmax}} \sum_{r=1}^R (-d(m_r, m_{r+1}|\boldsymbol{\theta}_r) + a(\bar{\tau}_r, \bar{\tau}_{r+1})) \end{aligned} \quad (10)$$

where

$$\begin{aligned} & d(S_r, S_{r+1}|\boldsymbol{\theta}_r) \\ &= \frac{1}{2} \sum_{k=S_r}^{S_{r+1}-1} \left[\log(2\pi\sigma^2) + \frac{(\log \tau_k - \log \tau(s_k|\boldsymbol{\theta}_r))^2}{\sigma^2} \right] \\ & a(\bar{\tau}_r, \bar{\tau}_{r+1}) = \log \left(1 - \exp\left(-\frac{(\bar{\tau}_r - \bar{\tau}_{r+1})^2}{2\sigma^2}\right) \right) \end{aligned}$$

and R is the number of sudden tempo alternations and is also the variable used in estimation. The r -th tempo curve $\tau(t|\boldsymbol{\theta}_r)$ is defined only in the range of $t_{S_r} \leq t < t_{S_{r+1}}$.

3.5. Optimization Algorithm of Tempo Model

Optimization of the model expressed by Eq. (10) can be achieved using the segmental k -means algorithm [2]. After the initial boundary is given, this algorithm is performed by iterating 2 steps: optimization and segmentation (see Fig. 9).

Optimization Step

Parameters of each rhythm pattern can be optimized by minimizing $d(m_r, m_{r+1}-1)$. Since this function is convex for the function $\tau(t)$, minimization can be formulated

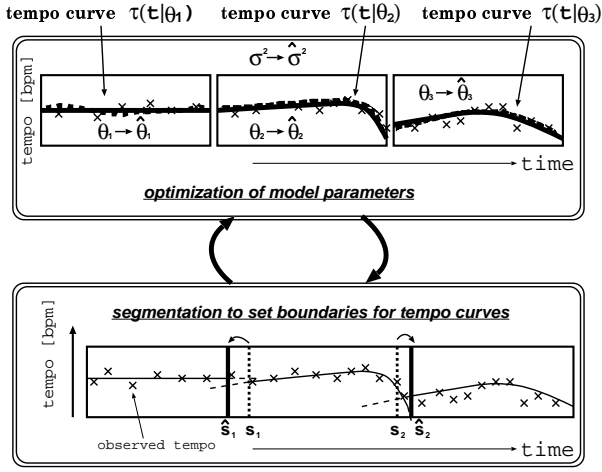


Figure 9. Iteration of segmentation and curve fitting in the segmental k -means algorithm. (conceptual diagram)

based on variable principle and carried out by setting the functional derivative $\delta d(m_r, m_{r+1}-1)$ to be 0.

$$\sum_{k=S_r}^{S_{r+1}-1} (\log \tau_k - \log \tau(t_k | \theta_r)) \cdot \delta \log \tau(t_k) = 0$$

From this, the optimal parameters of the model polynomial (Eq. (8)) are found by solving the following $P + 1$ equations:

$$\sum_{k=S_r}^{S_{r+1}-1} \log \tau_k \cdot t_k^{p'} - \sum_{k=S_r}^{S_{r+1}-1} \sum_{p=0}^P t_k^{(p+p')} \cdot a_p = 0$$

where $p' = 0, 1, \dots, P$.

Variance σ^2 in Eq. (7) is also optimized for all samples in the observed local tempo data with

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{r=1}^R \sum_{k=S_r}^{S_{r+1}-1} (\log \tau_k - \log \tau(t_k | \hat{\theta}_r))^2$$

In this optimization step, the parameters are updated for each of tempo curves $\{\theta_r\}_{r=1}^R$ and the variance of the tempo deviation σ^2 .

Segmentation Step

Boundaries of the segmented region of the tempo curve can be found efficiently using DP (Dynamic Programming) algorithm to maximize the objective function. We denote the cumulative log likelihood of m -th measure in the r -th tempo curve by $\delta_r(m)$, the number of measures by M , and the order of each measure by m . The algorithm is:

$$\begin{aligned} & r=1 \\ & \text{for } m=1 \text{ to } M \\ & \quad \delta_0(m) = d(0, m) \\ & \text{for } r=2 \text{ to } R \end{aligned}$$

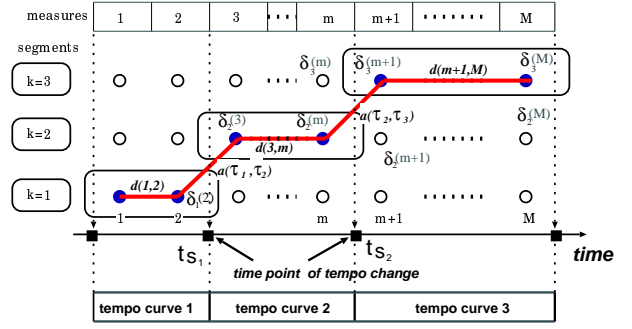


Figure 10. Dynamic Programming (DP) to detect bar lines at tempo changes.

for $m=k+1$ to M

$$\delta_r(m) = \max_{m' \in (k, \dots, m-1)} [\delta_{r-1}(m') + d(m', m) + a(\tau_r^-, \bar{\tau}_r)]$$

Here, the last node $\delta_p(M)$ gives the logarithm of the MAP probability, and the optimal path is obtained by trace-back. The most likely boundary is given by the path in the node trellis as shown in Fig. 10.

The number of tempo changes R is estimated with the MAP estimator of Eq. (10) by comparing the MAP probabilities for $R = 1, 2, \dots$

3.6. Experimental Example

A musical performance with an electronic piano recorded in SMF was modeled by tempo curves using the proposed model. To demonstrate the algorithm, we used “Fürchtenmachen”³ as an example with suddenly altering tempo between “Schneller(faster)” and original slow tempo several times within the piece.

Two kinds of tempo curves were tested on the performance of “Fürchtenmachen.” First, using a quadratic tempo curve model: $\log \tau(t) = a_0 + a_1 t + a_2 t^2$, the timings of tempo change were correctly estimated as shown in Fig. 11. Next, by fitting linear tempo curves: $\log \tau = a_0 + a_1 t$, detailed tempo behavior was extracted. In the MIDI recording of piano performance of “Fürchtenmachen,” the number of tempo changing time points and the locations of changing bar lines are estimated correctly.

The proposed method was also evaluated in estimation of the number of tempo changes and the bar-line locations at tempo-changing timings. The results were verified with MIDI data associated with the RWC music database of classical music [11] which had been manually prepared to approximately label the audio recording. Other experimental evaluation were also successful in RWC-MDB-C-2001, No. 1, Haydn’s “Symphony No. 94 in G major, Hob. I-94 ‘The Surprise’, 1st mvmt.”, and RWC-MDB-C-2001 No. 13, Mozart’s “String Quartet” No.19 in C major, K.465, 1st mvmt.

³ A piano piece from “Kinderszenen”, Op. 15, No.11, composed by Robert Schumann.

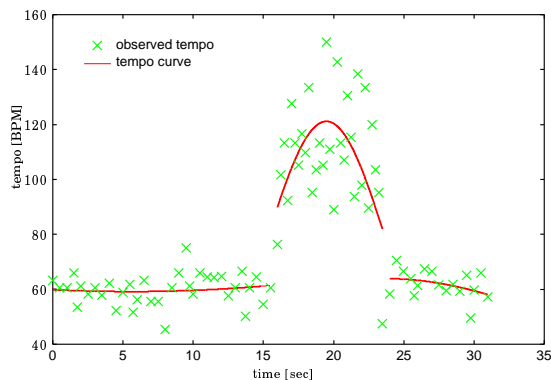


Figure 11. Example of quadratic tempo model ($\log \tau = a_0 + a_1 t + a_2 t^2$) fit to real performance: tempo-changing timings are detected correctly.

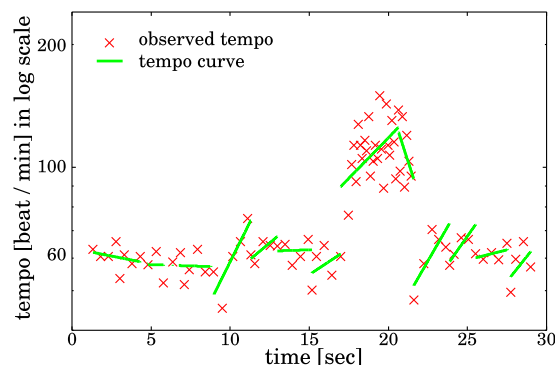


Figure 12. Example of linear tempo model ($\log \tau = a_0 + a_1 t$) fit to real performance: intra-phrase tempo changes are observed.

4. CONCLUSION

We have discussed rhythm recognition and tempo analysis of expressive musical performances, based on a probabilistic approach. Given a sequence of note durations deviated from nominal note lengths in the score, the most likely note values intended by the performer are found with the same framework as continuous speech recognition. This framework consists of stochastically deviating note durations modeled by HMMs and a stochastic grammar of “rhythm vocabulary” expressed with N -gram grammar. The maximum *a posteriori* note sequence is obtained by an efficient search using the Viterbi and level building algorithms. Significant improvements have been demonstrated compared with conventional “quantization” techniques. Tempo analysis is performed by fitting a parametric tempo curves to the observed local tempos for the purpose of extracting tempo dynamics and characteristics of the performance. Timings of tempo changes and optimal tempo curve parameters are simultaneously estimated using segmental k -means algorithm.

Future work includes integrating direct modeling poly-rhythm patterns, which includes synchronized multi-rhythm patterns, to give a direct relation between prob-

abilistic models and score data. Validity of the model should also be examined using audio recordings of professional instrumental players.

5. ACKNOWLEDGEMENTS

We thank Chandra Kant Raut for his valuable comments on English expressions in this paper.

6. REFERENCES

- [1] S. Sagayama, K. Takahashi, H. Kameoka, T. Nishimoto, “Specmurt Anasyllis: “A Piano-Roll-Visualization of Polyphonic Music Signal by Deconvolution of Log-Frequency Spectrum,” Proc. ISCA. SAPA, 2004, to appear.
- [2] L. Rabiner, B.-H. Juang: Fundamentals of Speech Recognition, Prentice-Hall, 1993.
- [3] N. Saitou, M. Nakai, H. Shimodaira, S. Sagayama, “Hidden Markov Model for Restoration of Musical Note Sequence from the Performance,” Proc. of Joint Conf of Hokuriku Chapters of Institutes of Electrical Engineers, Japan, 1999, F-62, p.362, Oct 1999. (in Japanese)
- [4] N. Saitou, M. Nakai, H. Shimodaira, S. Sagayama: “Hidden Markov Model for Restoration of Musical Note Sequence from the Performance,” Technical Reports of Special Interest Group on Music and Computer, IPSJ, pp. 27–32, 1999. (in Japanese)
- [5] A. Cemgil, B. Kappen, P. Desain, H. Honing: “On tempo tracking: Tempogram Representation and Kalman filtering,” J. New Music Research, vol. 29, no. 4, 2000.
- [6] C. Raphael: “Automated Rhythm Transcription,” Proc. ISMIR, pp. 99–107, 2001.
- [7] P. Trilsbeek, P. Desain, H. Honing: “Spectral Analysis of Timing Profiles of Piano Performances,” Proc. ICMC, 2001.
- [8] S. Dixon, W. Goebel and G. Widmer The Performance Worm: “Real Time Visualisation of Expression Based on Langner’s Tempo-Loudness Animation,” Proc. ICMC, pp 361-364. 2002.
- [9] L. R. Rabiner, B. H. Juang: “An Introduction to Hidden Markov Models,” IEEE ASSP magazine, pp. 4–16, 1986.
- [10] C. Myers, L. R. Rabiner: “Connected Digit Recognition Using Level-Building DTW Algorithm,” IEEE Trans. ASSP, Vol. 29, pp. 351–363, 1981.
- [11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka: “RWC Music Database: Popular, Classical, and Jazz Music Databases,” Proc. ISMIR, pp.287-288, 2002.