RHYTHM AND TEMPO ANALYSIS TOWARD AUTOMATIC MUSIC TRANSCRIPTION

Haruto Takeda, Takuya Nishimoto, Shigeki Sagayama

Graduate School of Information Science and Technology
The University of Tokyo
Bunkyo-ku, Tokyo, 113-8656, Japan

ABSTRACT

This paper discusses model-based rhythm and tempo analysis of music data in the MIDI format. The data is assumed to be obtained from a module performing multi-pitch analysis of music acoustic signals inside an automatic transcription system. In performed music, observed note lengths and local tempo fluctuate from the nominal note lengths and long-term tempo. Applying the framework of continuous speech recognition to rhythm recognition, we take a probabilistic top-down approach on the joint estimation of rhythm and tempo from the performed onset events in MIDI data. Short-term rhythm patterns are extracted from existing music samples and form a "rhythm vocabulary." Local tempo is represented by a smooth curve. The entire problem is formulated as an integrated optimization problem to maximize a posterior probability, which can be solved by an iterative algorithm which alternately estimates rhythm and tempo. Evaluation of the algorithm through various experiments is also presented.

Index Terms— Rhythm recognition, rhythm vocabulary, rhythm N-gram model, piecewise polynomial tempo curve, HMM, Viterbi search

1. INTRODUCTION

Automatic music transcription (AMT) has long been one of the ultimate goals of music information processing. Just like automatic speech recognition (ASR) which converts speech signals to text, AMT converts either audio signals or MIDI (Musical Instrument Digital Interface) data to music score. A reconstructed music score would have many applications such as music visualization, sheet music publishing, music input to computers followed by automatic arrangements, or music information retrieval both for music database construction and query input. This paper discusses MIDI-to-score conversion, which is to follow audio-to-MIDI conversion [1] inside a newly integrated music transcription system [2]. Since pitch is already given in the MIDI data, we focus on the joint estimation of rhythm and tempo, which takes as input the timing information of the performed notes represented as MIDI events, and outputs a symbolic rhythm expression for the corresponding musical score.

In music, each music note has a note value (nominal length of the notes in the score) described in the source score. In performed music, however, observed note-on and note-off MIDI events deviate from the timings expected from the note values and over-all tempo. Thus, the inverse problem of going back from the performed music to the source score is not a trivial one, in the same way as the speech-to-text conversion problem.

Rhythm recognition has been regarded in the past as the integration of bottom-up processings: first, *beat induction* finds the beat intervals and *beat tracking* tracks the beat onsets, usually in real time. Then, for each beat, *quantization* is performed to convert the time length of each note into note values, and *meter analysis* is performed to find the meter structure and estimate the time signature.? Each of these tasks was studied from rule-based approaches [3, 4], and some of /Rtkdthe tasksthem were combined to extract automatically beat and meter from musical performances [5]. Recently, quantization and beat tracking were also integrated in probabilistic top-down approaches [6, 7], where time develoepment of onset times and tempo were probabilistically modeled. These methods assumed that time signature was given *a priori*, while tempo and onset times were modeled in the two consecutive notes.A MAP (maximum *a posteriori* probability) estimation approach was also introduced, relying on approximation of integral by random sampling method [7] and dynamic programming (DP) [6]. While these tasks were performed on onset timing information, some of them can be preceded by audio preprocessing including onset detection and frequency analysis [5, 8], even though onset detection error is inevitable.

On the other hand, since 1999 we followed a top-down HMM-based approach [9, 10] motivated by the ASR-like idea that rhythm should be 'recognized' as a pattern, without relying on quantization and tracking, and modeled using vocabulary and grammar. We also introduced the empirical constraint that tempo usually changes smoothly with a continuous function in the same way as it is dealt with in music performance analysis [11, 12, 13]. This paper presents probabilistic models of rhythm vocabulary and note-length fluctuations which are combined together to find the most likely rhythm and tempo in a similar manner as is done in continuous speech recognition (CSR) [14] where language and acoustic models work cooperatively. One of the points which distinguishes our

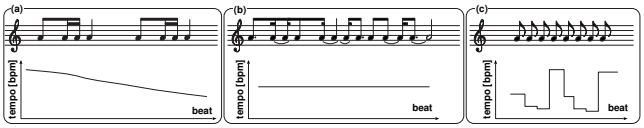

Fig. 1. Music transcription is essentially an ill-posed problem: while rhythm (a) is intended with the tempo slowing down, the resulted note onset timings can be interpreted both as (b) assuming a constant tempo, and even as (c) allowing rapid changes in tempo.

work from previous efforts is the introduction of a *grammar* model for rhythm. The use of HMMs including tempo as a latent variable also differentiates the estimation algorithm of the rhythm recognition from that of CSR.

2. MODELING RHYTHM AND TEMPO

2.1. Addressing the problem

In music performances, the observed note length x [secs] is basically the nominal length ('note value') q [beats] multiplied by the tempo r [seconds per beat, in this paper] with additional deviations caused by artistic intentions and/or insufficient playing skills. Therefore,music transcription is an ill-posed inverse problem, which consists into determining both rhythm and tempo simultaneously, i.e. to reconstruct the original (or intended) score, and has no unique solution in general as seen in Fig. 1.

However, humans listening to the performed music piece shown in Fig. 1 generally prefer score (a) rather than (b) and (c). This preference can be explained by two hypotheses: (1) humans recognize a rhythm as following a simple template from a collection of common rhythms, allowing small deviations of note onset timings, and (2) tempo changes only smoothly within short periods. These hypotheses are considered to be commonly accepted from the viewpoint of the nature of music performances.

2.2. Modeling note values

Note values are discrete quantities rational to beat, e.g. a whole note is 4 beats assuming the quarter note as the beat unit. For a monophonic music piece consisting of N notes, its rhythm score can be expressed as $Q = \{q_1, \cdots, q_N\}$, where q_n [beats] represents the note value of the n-th note.

We define here the concepts of "rhythm words" and "rhythm vocabularies". A note value sequence of relatively short length is considered as a rhythm word w_j, and the set of all these rhythm words constitutes a rhythm vocabulary, just like a word vocabulary in CSR. The rhythm score Q is then considered as a sequence of rhythm words and is associated with a *prior probability* $P(Q)$ representing how likely the rhythm Q is to appear in the source score. This introduces a stochastic grammar on the rhythm.

As in CSR, $P(Q)$ can be approximated by an N-gram probability on the rhythm words, which are considered as short range rhythm pattern units. If meters are chosen as separators between rhythm words, estimation of the rhythm words

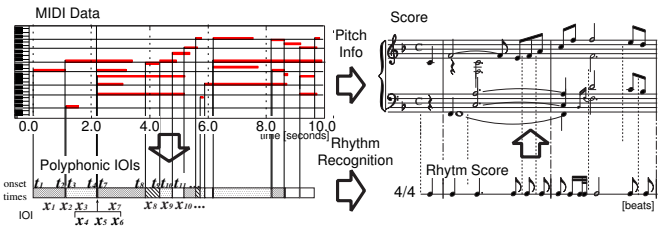

Fig. 2. Polyphonic music transcription: MIDI data is converted to a single stream of IOIs, recognized as a sequence of likely note onsets and finally converted to score using pitch information.

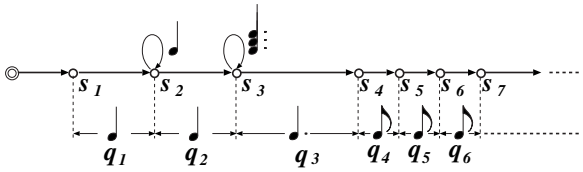

Fig. 3. Modeling rhythm score with Markov transition of beat position.

also works as meter analysis since bar line positions are then obtained as their boundaries, and time signature is obtained from the most likely rhythm words. The rhythm vocabulary and its N-gram grammar can be trained using original scores of composed music samples. The problem of stochastic training with a limited amount of data can be discussed in the same way as for language modeling in speech recognition with the out-of-vocabulary problem.

Estimation of polyphonic rhythm can be accomplished in the same way as in the monophonic case by projecting all observed note onsets onto a one-dimensional time axis and by using an inter-voice rhythm vocabulary to recognize the note values in the one-dimensional rhythm score from which the source score is estimated using pitch information in the MIDI data as shown in the right of Fig. 2. The rhythm vocabulary of polyphonic music can be trained in the same way as in the monophonic case on a set of polyphonic music scores.

The beat position of the n-th note in a score is represented by the cumulative note values $s_n = \sum_{i=1}^{n-1} q_i$. Regarding the score positions of all the notes in a rhythm score as nodes of a directed graph, a rhythm score can be modeled by a state transition network such as shown in Fig. 3. Chords can be modeled as self-transitions at a node in the same way as in previous works[6, 15].

2.3. Modeling tempo with a tempo curve

We denote by $R(s)$ [sec/beat] the tempo (meaning time length per beat, in this paper) played at beat position s in the score. Observed instantaneous tempo (observed note length divided by its note value) at position s is modeled as a sample of $R(s)$ including deviations (statistical errors) caused by musical intentions, insufficient playing skills, etc. Rapid or sudden changes in tempo can be modeled by switching tempo curves, where tempo curves are not continuous at the points of these tempo changes.

In the following discussion, we use segmental polynomi-

als in log-scaled tempo to model the tempo curves. We define in each segment $\log R^{(k)}(s) = \sum_{d=0}^{D} a_d^{(k)} s^d$, where k is the segment index, $a_d^{(k)}$ denotes the d-th coefficient of the polynomial in the k-th segment and D the maximal polynomial order.

2.4. Note length and HMM

Note lengths of polyphonic music can be obtained as inter-onset intervals (IOIs) as shown in Fig. 2, which correspond to the observation of rhythm score.

As already stated, the note length x_n of the n-th note played at beat position s_n is expected to be the note value q_n multiplied by the tempo $R(s_n)$ plus a deviation. Since the deviation of the note duration is expected to be roughly proportional to the note value, we assume that the distribution the logarithmic note length $\log x$ follows a normal distribution with mean $\log(q_n \cdot R(s_n))$ and variance σ^2. As the probabilistic distance between x and $q_n \cdot R(s_n)$ is given in the logarithmic scale, it can also be understood as the equivalent distance for the tempo by noticing that

$$\log x_n - \log(q_n \cdot R(s_n)) = \log r_n - \log R(s_n), \quad (1)$$

where r_n is the instantaneous tempo of the n-th note defined by $x_n$secs $= r_n$secs/beat $\cdot\, q_n$beats.

Note that we are using the same probability distribution $P(x|q, R(s))$ for both IOI and tempo, while previous works introduced a probability for each of the IOI and the tempo.

Although all notes in a chord are supposed to be played simultaneously, their onset timings are not exactly the same in real performances. We assume that IOIs between notes in a chord follow a single-sided Gaussian distribution.

Combining together the beat position transition Markov model and the probabilistic model of the IOIs, the process of generating IOIs from a sequence of rhythm words is modeled using HMMs which expressthe probability $P(X|Q, R)$ of observing IOIs X such that rhythm Q in the score is played at tempo R.

2.5. Probability of generating a music performance

Combining further these models, the probability that the rhythm score Q produces a sequence of IOIs X with tempo curve $R(s)$ is given by $P(X|Q, R)P(Q)P(R)$. This can be understood as a stochastic process which consists in emitting IOIs at the transitions between the states (beat positions) of an HMM network where the HMMs (rhythm words) are probabilistically connected through an N-gram grammar.

3. JOINT ESTIMATION OF RHYTHM AND TEMPO

3.1. Maximum a posteriori probability estimation

Joint estimation of note values and tempo can be formulated as an estimation of the note values sequence Q and tempo curve $R(s)$ from the IOI sequence $X = \{x_1, x_2, \cdots, x_N\}$ of a real performance. Assuming that Q and R are independent, the posterior probability can be written as

$$P(Q, R|X) \propto P(X|Q, R)P(Q)P(R) \quad (2)$$

according to Bayes' rule. Though it is not a trivial problem to find the combination of Q and R that maximizes Eq. 2, iterating the alternate estimation of the note value sequence Q and the tempo R monotonically increases the posterior probability, which thus converges to a (locally) optimal solution with regards to both rhythm and tempo. The optimal solution of the MAP estimation can be obtained by using an appropriate initial condition.

3.2. Algorithm for rhythm recognition

With the tempo curve $R(s)$ fixed, the most likely rhythm Q for given IOI X can be estimated. Viewed from the analogy between speech recognition and rhythm recognition, this is formulated as a best path-search problem in a state transition network consisting of HMMs. An efficient search algorithm based on one-pass DP (e.g. time synchronous Viterbi search) can be used for this problem.

3.3. Algorithm for tempo curve estimation

On the other hand, with the note value sequence Q fixed, the posterior probability can be monotonically increased by re-estimating the tempo curve $R(s)$. Maximizing the logarithmic posterior probability of the tempo curve is equivalent to the least squares estimation, assuming that no *a priori* knowledge $P(R)$ is given, i.e. that $P(R)$ is uniform. Fitting segmental polynomials can also be done using the segmental k-means method[16] by iteratively and monotonically increasing the posterior probability.

3.4. Procedure to jointly estimate the rhythm and tempo

Combining rhythm recognition and tempo estimation discussed above, the joint estimation of rhythm and tempo can be accomplished as follows:

1. **Extract IOIs** from MIDI events in the given performed music.
2. **Set initial condition of tempo curve** $R(s)$**.** (e.g. constant tempo R_c when *a priori* knowledge is not available)
3. **Rhythm recognition**: Assuming the tempo curve as $R(s)$, find Q that maximizes the posterior probability for the given X by using the Viterbi algorithm, performing an optimal path-search in the rhythm vocabulary HMMs.
4. **Tempo estimation**: Fixing the estimated Q, re-estimate the tempo curve $R(s)$ that maximizes the posterior probability for the instantaneous tempo X/R using the segmental k-means algorithm.
5. **Convergence test**: Terminate if the increase of the posterior probability is less than a preset threshold. Otherwise, go back to (3).

This iterative algorithm is substantially the process of minimizing the probabilistic distance between x_n and $q_n \cdot R(s_n)$ given by $p(x_n|q_n, R(s_n))$ within the rhythm vocabulary and tempo curve constraints.

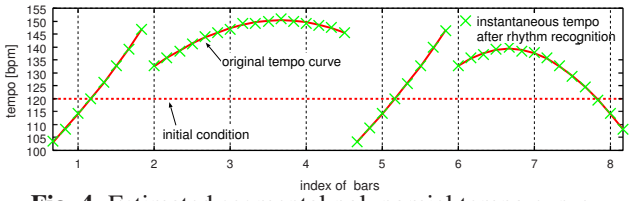

Fig. 4. Estimated segmental polynomial tempo curve.

4. EXPERIMENTAL EVALUATION

4.1. Evaluation of the tempo curve estimation

First, we conducted a simulated test of the algorithm to estimate the tempo curve from MIDI data[1] generated from artificially given tempo curves so that the played MIDI data sounded well like a human performance. Setting the initial constant tempo to 120 bpm and using a bigram model rhythm vocabulary trained with 133 pieces, 4 second-order segmental polynomials were iteratively fit to the observed instantaneous tempos to log-normal distribution with variance of 0.1 until the convergence test cleared a threshold of 0.0001. As shown in Fig. 4, the estimated tempo curve exactly fits the original tempo curve after 2 iterations.

4.2. Evaluation of the note values estimation

Next, to confirm the performance for transcription, we evaluated the rhythm recognition accuracy of the proposed method over 37 piano pieces[2] played by 2 piano players with an electronic piano and recorded in the MIDI format. Two kinds of rhythm vocabularies ('open' and 'closed') and their bigram models were trained using 100 piano pieces (distinct from the pieces used for testing) for the open vocabulary and 137 pieces for the closed vocabulary (including the pieces used for testing, which corresponds to assuming that there is no out-of-vocabulary words), adding to the open vocabulary the 37 pieces used in the evaluation. The closed model was prepared to evaluate the rhythm recognition performance under a condition free from out-of-vocabulary rhythm words.

Note values and chord clustering of the estimated rhythm score were compared with those of the original scores. A rhythm accuracy of 85.5% was attained with the 'closed' vocabulary, 81.9% with the 'open' vocabulary. Some errors occurred when note values were misrecognized in parts of a piece where the same note value is repeated several times (e.g. a sequence of triplets and that of eighth notes), or when tempo curve estimation fell into local optimums.

4.3. Reconstructing the score

Combining the pitch information of the input MIDI data and the estimated rhythm score, a music score can be reconstructed. The key signature, unit beat length and accidentals are yet to be determined automatically, and the key was given manually while the beat length and accidentals were derived by an ad hoc method this time. Using a closed vocabulary for the rhythm words, we obtained a score shown in Fig. 5 that can be considered essentially the same as the original score.

[1] 8 bars from Bagatelle "Für Elise," WoO. 59 by L. van Beethoven.

[2] Etudes, Op. 100 by F. Burgmüller, Kinderszenen, Op. 15 by R. Schumann, and 4 Mazurkas by F. Chopin.

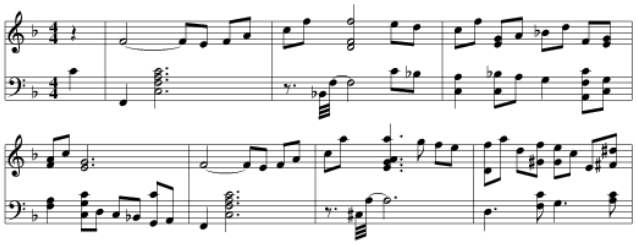

Fig. 5. An example of reconstructed score from a piano performance of "Träumerei" by R. Schumann, Op. 15-7.

5. CONCLUSION

This paper discussed the joint estimation of rhythm and tempo of the MIDI data of a human performed music piece. Future work includes cross-voice rhythm modeling for voice description, and integration of pitch aspects including key finding and chord analysis toward automatic transcription.

This research was partly supported by MEXT Grant-in-Aid #17300054 and CrestMuse Project of JST.

6. REFERENCES

[1] H. Kameoka, *et al.*,"Harmonic-temporal structured clustering via deterministic annealing EM algorithm for audio feature extraction," *Proc. ISMIR*, pp. 115-122, 2005.

[2] K. Miyamoto, *et al.*, "Probabilistic approach to automatic transcription from audio signals," submitted to *ICASSP2007*, 2007.

[3] D. Rosenthal, "Emulation of Human Rhythm Perception," CMJ 16(10), pp. 64-76, 1992.

[4] D. Temperley, Cognition of Basic Music Structure, MIT Press, 2001.

[5] S. Dixon, "Automatic Extraction of Tempo and Beat from Expressive Performances," JNMR 30(1), pp. 39-58, 2001.

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

[7] A. T. Cemgil *et al.*, "Monte Carlo Methods for Tempo Tracking and Rhythm Quantization," JAIR, 18(4), pp. 45–81, 2003.

[8] A. Klapuri *et al.*, "Analysis of the Meter of Acoustic Musical Signals," IEEE Trans. ASLP-14(1), 2006.

[9] N. Saito, *et al.*, "Hidden Markov Model for Restoration of Musical Note Sequence from the Performance," *Proc. the Joint Conference of Hokuriku Chapters of Institutes of Electrical Engineers*, F-62, pp. 362, 1999. (in japanese)

[10] H. Takeda *et al.*, "Hidden Markov Model for Automatic Transcription of MIDI Signals," *Proc. MMSP2002*, 2002.

[11] N. P. M. Todd, "The Dynamics of Dynamics: A Model of Musical Expression," JASA-91(6), pp. 3540-3550, 1992.

[12] B. H. Repp, "Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's Träumerei," JASA, 92(5), 1992.

[13] H. Honing, "The Final Retard: On Music, Motion, and Kinematic Models," CMJ 27(3), pp. 66-72, 2003.

[14] X. Huang, *et al.*, Speech Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001.

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

[16] B. H. Juang and L. R. Rabiner, "The Segmental K-Means Algorithm for Estimating Parameters of Hidden Markov Models," *IEEE Trans. ASSP-38(9),* pp. 1639-1641, 1990.