REAL-TIME AUDIO TO SCORE ALIGNMENT USING SEGMENTAL CONDITIONAL RANDOM FIELDS AND LINEAR DYNAMICAL SYSTEM

Ryuichi Yamamoto, Shinji Sako and Tadashi Kitarmura
Graduate School of Engineering, Nagoya Institute of Technology, Japan
{ryuichi, sako, kitamura}@mmsp.nitech.ac.jp

ABSTRACT

This extended abstract describes our polyphonic score follower submitted to MIREX 2012 Real-time Audio to Score Alignment (a.k.a score following) task. Our method employs Segmental Conditional Random Fields (SCRFs) and Linear Dynamical System (LDS) for modeling discrete beat transitions and continuous tempo fluctuation. In the decoding process, delayed decision viterbi algorithm leads a robust beat estimation and kalman filtering algorithm leads a very fast estimation of time-varying tempo. The continuous beat position at the current time is anticipated using the results of these two algorithms.

1. SYSTEM OVERVIEW

An overview of our system is shown on Figure 1. The input is the incoming audio signal divided into overlapping time frames and the output is the continuous beat position on the score. The score corresponding the audio is given in advance as a standard MIDI file. In the feature extraction process, the chroma vector and the onset feature are extracted for each frame, that represent pitch information and whether a onset is detected or not by gabor wavelet transform and the onset detector based on spectral flux introduced by Dixon [1], respectively.

Our real-time audio to score alignment algorithm consists of three steps: (1) discrete beat (chord) and its inter-onset-interval (IOI) estimation using delayed decision viterbi algorithm, where the chord denotes concurrent notes on the score, (2) time-varying tempo estimation using kalman filtering algorithm, (3) continuous beat anticipation using the result of these two algorithms (see section 4 for details).

2. SEGMENTAL CRF FOR SCORE ALIGNMENT

2.1 Segmental CRF formulation

We describe the audio to score alignment problem as the segmentation of the audio to the chord sequence on the score. In our framework, discrete beat transitions between chords are modeled by SCRFs, that are extension of Conditional Random Fields (CRFs) which markovian assumption is relaxed to segment level from frame level. CRFs and SCRFs have been first introduced to the audio to score alignment problem by Joder [2]. They can allow more flexible feature design than conventional Hidden Markov Models (HMMs). In particular, SCRFs can incorporate frame level features but also segment level features.

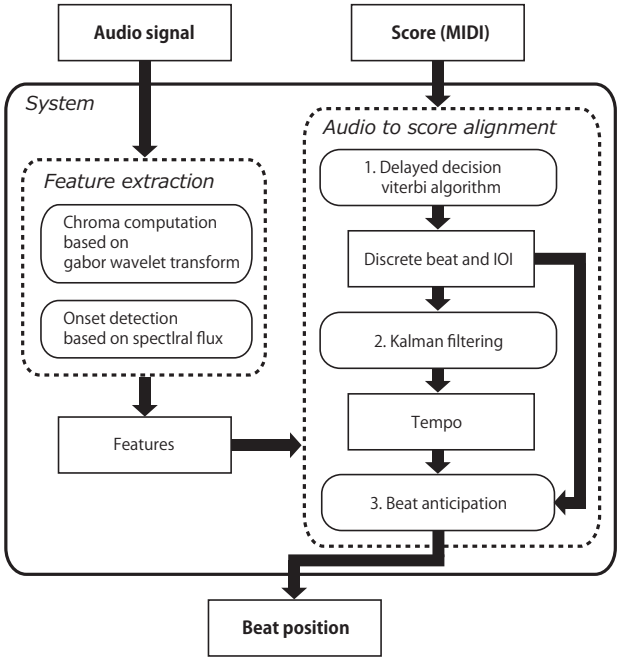

Figure 1. Overview of our real-time audio to score alignment system submitted to MIREX 2012.

Let $\mathbf{o} = \{\boldsymbol{o}_t\}_t$ be the observation sequence extracted from the input audio signal where t is the frame index, and let $\mathbf{q} = \{q_n\}_n$ be the segmentation of $\mathbf{o}$, where n is the segment index, the segment $q_n = (t_n^s, t_n^e, s_n)$ consists of the start frame t_n^s , the end frame t_n^e, and the chord state s_n, respectively. The alignment problem is formulated as

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathcal{Q}} p(\mathbf{q}|\mathbf{o}), \tag{1}$$

where $\mathcal{Q}$ is a set of possible segmentations. The conditional probability given the observation sequence is defined as

$$p(\mathbf{q}|\mathbf{o}) = \frac{1}{Z}\exp\Big\{\sum_{n,k}\lambda_k^t f_k^t(q_n, q_{n-1}) + \sum_{n,l}\lambda_l^o f_l^o(q_n, \mathbf{o}))\Big\}, \tag{2}$$

where Z is the normalization factor over segmentations,

$f_k^t(q_n, q_{n-1})$ is a transition feature, $f_l^o(q_n, \mathbf{o})$ is an observation feature, λ_k^t and λ_l^o are the feature weights that are the parameters of the model, respectively. The most likely segmentation is estimated using viterbi algorithm.

2.2 Features

2.2.1 Transition features

Transition features that used in our SCRF framework are as follows:

- log transition probability of HMMs
- the length of a segment

Performance repetitions and jumps are modeled as the features of log transition probability of HMMs. The feature of a segment length represents the duration of the concurrent notes.

2.2.2 Observation features

Observation features based on the observed chroma vector and the onset feature are as follows:

- Kullback-Leibler (KL) divergence between the observed chroma and the template chroma build from the score.
- the top of a segment is onset or not
- the number of onsets is detected in a segment

The KL divergence between the observed chroma and the template chroma represents a matching measure of pitch information. The chroma template is build in advance from the score for each chord, as in [3]. In order to incorporate the burst of concurrent notes, onset features are designed for each segment.

3. LDS FOR TEMPO ESTIMATION

Time-varying tempo during the performance is modeled as Linear Dynamical System (LDS). It is assumed that the tempo of the performance can fluctuate locally but can be considered smooth.

Let r_n be the local tempo (sec / beat) that is considered as constant in the segment, d_n the duration (sec), and let l_n be the chord length (beat) on the score. The tempo model is defined as

$$r_n = r_{n-1} + w_n, \quad (3)$$

$$d_n = r_n l_n + v_n, \quad (4)$$

where

$$w_n \sim \mathcal{N}(0, Q), v_n \sim \mathcal{N}(0, R), \quad (5)$$

$\mathcal{N}(\cdot)$ denotes the density function of Gaussian distribution. Q and R are the standard variance of the distributions. Given the result of chord segmentation, time-varying tempo is estimated using kalman filtering algorithm in an iterative manner.

4. REAL-TIME DECODING

4.1 On-line approximation

A simple greedy approximation that reports the most probable current state is often applied to viterbi algorithm or dynamic time warping for score following [4,5]. However, it may cause estimation errors especially in the input audio is complex. To avoid this problem, we use delayed decision viterbi algorithm for on-line approximation that report the most probable α-time past state.

Now we describe our beat anticipation algorithm. Let $\{\hat{s}_1, \ldots, \hat{s}_{t-\alpha}, \ldots, \hat{s}_t\}$ be the result of chord segmentation at time t, $\{\hat{r}_1^{-1}, \ldots, \hat{r}_{t-\alpha}^{-1}, \ldots, \hat{r}_t^{-1}\}$ the result of the reciprocal of tempo estimation, and let $\{b_1, \ldots, b_{t-\alpha}, \ldots, b_t\}$ be the beat sequence corresponding the estimated chord sequence. The current beat position at time t is anticipated as

$$\hat{b}_t = b_{t-\alpha} + \int_{t-\alpha}^{t} r_\tau^{-1} d\tau. \quad (6)$$

The above equation can be approximated as

$$\hat{b}_t = b_{t-\alpha} + r_{t-\alpha}^{-1} \cdot \alpha, \quad (7)$$

when we assume the tempo is considered as constant from $t - \alpha$ to the current time t. In our system, α is set to the time from the estimated previous segment to the current time (often less than 1 sec). Note that α is constant in segment level but variable in frame level.

Our real-time beat anticipation algorithm is summarized below.

Step 1 Chord segmentation using delayed decision viterbi algorithm for the input observation sequence

Step 2 Tempo estimation using kalman filtering algorithm given the result of chord segmentation

Step 3 Beat anticipation using the results of Step 1 and Step 2, where the delayed time α is set to the time from the estimated previous segment to the current time

4.2 Relaxation of computational complexity

Forward calculation of vierbi algorithm for each time frame have the computational cost of $\mathcal{O}(S^2)$ if we consider all possible chord transitions, where S is the number of chords on the score. It may be difficult to carry out in real-time when S is a large number (e.g. over 2000). Thus we introduce some assumptions as follows: (1) Chord transitions are restricted to the near the maximum a posteriori (MAP) chord on the previous time frame. (2) Repetitions and jumps are occurred at the specified points in advance (actually random in our setting). The transition probabilities to the chords except for these chords are set to zero and forward calculation can be omitted.

5. RESULT

The result of MIREX 2012 Real-time Audio to Score Alignment is shown on ...

6. ACKNOWLEDGEMENT

This research was supported in part by the Ichihara International scholarship foundation.

7. REFERENCES

[1] S. Dixon: “Onset detection revisited,” *Proc. Digital Audio Effects*, 2006.

[2] C. Joder, S. Essid and G. Richard: “A Conditional Random Field Framework for Robust and Scalable Audio-to-Score Matching,” *IEEE, Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2385–2397, 2011.

[3] N. Hu, R. B. Dannenberg and G. Tzanetakis: “Polyphonic Audio Matching and Alignment for Music Retrieval,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 185–188, 2003.

[4] N. Orio and F. Dechelle: “Score Following Using Spectral Analysis and Hidden Markov Models,” *Proc. of the International Computer Music Conference*, Havana, Cuba, 2001.

[5] K. Suzuki, Y. Ueda, S. A. Raczynski, N. Ono, and S. Sagayama: “Real-time Audio to Score Alignment Using Locally-constrained Dynamic Time Warping of Chromagrams” *MIREX submission*, 2011.