ROBUST ON-LINE ALGORITHM FOR REAL-TIME AUDIO-TO-SCORE ALIGNMENT BASED ON A DELAYED DECISION AND ANTICIPATION FRAMEWORK

Ryuichi Yamamoto, Shinji Sako, Tadashi Kitamura

Graduate School of Engineering, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan

ABSTRACT

In this paper, we present a robust on-line algorithm for real-time audio-to-score alignment based on a delayed decision and anticipation framework. We employ Segmental Conditional Random Fields and Linear Dynamical System to model musical performance. The combination of these models allows an efficient iterative decoding of score position and tempo. The combined advantages of our approach are the delayed-decision Viterbi algorithm which utilizes future information to determine past score position with high reliability, thus improving alignment accuracy, and the fact that the future position can be anticipated using an adaptively estimated tempo. Experiments using classical music and jazz databases demonstrate the validity of our approach.

Index Terms— Audio-to-score alignment, score following, real-time, segmental conditional random fields, linear dynamical system

1. INTRODUCTION

Real-time audio-to-score alignment involves synchronizing an audio performance and its symbolic musical score, known as score following. It can be used in a wide range of real-time applications, such as the synchronization of live sounds and automatic accompaniment of human soloists or singers (e.g., see [1, 2]).

Audio-to-score alignment can be considered as either an *off-line* or *on-line* problem. In off-line cases, global information about the input audio signal can be used in the alignment process. In contrast, in on-line cases, we cannot use future information about the input signal. For this reason, score following is fundamentally more difficult compared than the off-line problem.

In both settings, many current approaches use dynamic programming methods based on Hidden Markov Models (HMMs) or Dynamic Time Warping (DTW). In the off-line setting, the most likely alignment can be found using a dynamic programming technique, given the entire input audio. However, the decoding algorithm is off-line, so some approximations are required for on-line cases.

In [3–5], a greedy approximation was applied to Viterbi algorithm and dynamic programming. In the case of polyphonic music, however, the number of estimation errors will increase. These are caused by uncertainties in pitch and onset, which increase in proportion to the complexity of the input audio. Thus, the greedy solution may not always be suitable.

In contrast to the dynamic programming approach, filtering methods based on state-space models have been proposed [6–8]. Although these allow the simultaneous estimation of score position and tempo, they tend to accumulate errors and fail to recover if they lose their position.

To maintain robustness against polyphonic music signal, it is helpful to use the tempo for future anticipation. In [9], Raphael used hybrid graphical models for the score position and tempo, but this technique only works off-line. In [10], Cont used duration-focused models consisting of Hidden Markov/Semi-Markov Models with an explicit tempo model. Behind the success of this work, the greedy approximation in the Viterbi algorithm may cause estimation errors with highly polyphonic signals. In [11], Arzt reported a sophisticated on-line algorithm, utilizing a forward-backward strategy that re-computes past-determined forward path, albeit without an explicit tempo model.

In this work, we introduce a robust on-line algorithm for polyphonic music signals based on a delayed decision and anticipation framework. The advantages of our approach are that a delayed decision approximation for the Viterbi algorithm can find highly reliable past positions utilizing future information, even in polyphonic cases, and future position can be anticipated using an adaptively estimated tempo. In addition, we employ the state-of-the-art, Segmental Conditional Random Fields (SCRFs) proposed in [12] (with a few modifications) and an explicit tempo model based on Linear Dynamical System (LDS).

2. SEGMENTAL CONDITIONAL RANDOM FIELDS OF MUSICAL PERFORMANCE

2.1. Score Alignment Formulation

We first describe the audio-to-score alignment problem in the off-line situation because score following can be approximated from its on-line extension.

Given the auditory music signal and its symbolic score, we address the score alignment problem as the segmentation of the audio to the chord sequence on the score (Figure 1), where a chord is a set of concurrent notes on the score. In our approach, chord transitions are modeled by SCRFs. SCRFs are an extension of Conditional Random Fields (CRFs) which Markovian assumption is relaxed to allow a segment-level

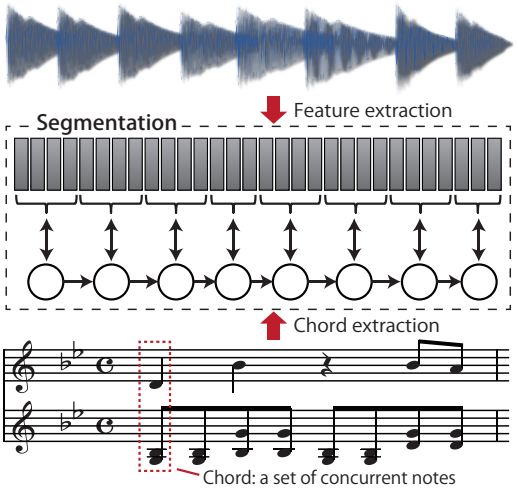

Fig. 1. Audio-to-score alignment as segmentation into a chord sequence from the feature vectors that are extracted by the input audio signal.

that is separate from the frame-level. CRFs and SCRFs were first introduced to audio-to-score alignment by Joder [12, 13]. They allow more flexible feature design than conventional HMMs. In particular, SCRFs can incorporate both frame-level and segment-level features.

In contrast to Joder's previous work, we model time-varying tempo as a continuous process rather than as a discrete process, which is discussed in Section 3. This allows for an adaptive real-time estimation of tempo. Note that tempo is not considered here as we simply allow inference using the Viterbi algorithm.

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 label s_n. The segmentation 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 the set of possible segmentations. The conditional probability of a given observation sequence is defined as

$$p(\mathbf{q}|\mathbf{o}) = \frac{1}{Z(\mathbf{o})}\Psi(q_1)\prod_{n=2}^{N}\Psi(q_{n-1}, q_n)\prod_{n=1}^{N}\Phi(q_n, \mathbf{o}), \tag{2}$$

where N is the number of segments, $Z(\mathbf{o})$ is a normalization factor, $\Psi(q_n, q_{n-1})$ are the transition functions, and $\Phi(q_n, \mathbf{o})$ are the observation functions. Indeed, N is a random variable. The most likely segmentation can be found using the Viterbi algorithm in the off-line setting thanks to the segment-level Markovian assumption.

2.2. Observation Functions

For the observation functions, which represent the relation between observations and chords, we use two acoustic features: chroma features based on a constant-Q transform to utilize the pitch content of musical performance , and onset features based on spectral flux to consider the burst-of-note onset. An extensive study on acoustic features in score alignment can be found in [14] and we follow this.

Let $\boldsymbol{o_t} = \{\boldsymbol{v_t}, f_t\}$ be the observation extracted from the audio signal, where $\boldsymbol{v_t}$ is the chroma vector and f_t is the result of the spectral flux-based onset detection proposed in [15].We assume that the observation functions can be decomposed as

$$\Phi(q_n, \mathbf{o}) = \phi_c(q_n, \boldsymbol{v}_{t_n^s:t_n^e})\phi_a(q_n, f_{t_n^s:t_n^e}). \tag{3}$$

a) Chroma feature: For each segment, a chroma feature is calculated as

$$\phi_c(q_n, \boldsymbol{v}_{t_n^s:t_n^e}) = \exp\Big\{-\lambda^c \sum_{t=t_n^s}^{t_n^e} D^{\mathrm{KL}}(\boldsymbol{v}_t||\boldsymbol{u}_{s_n})\Big\}, \tag{4}$$

where λ^c is a weighting parameter, $D^{\mathrm{KL}}(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence, and $\boldsymbol{u}_{s_n}$ is a template chroma vector built from the score for each chord in the same manner as described in [16].

b) Onset feature: An onset feature is defined as

$$\phi_a(q_n, f_{t_n^s:t_n^e}) = \exp\Big\{\sum_{g=0}^{1}\lambda_g^a\delta_{\{f_{t_n^s},g\}} + \sum_{h}\mu_h^a\delta_{\{m,h\}}\Big\}, \tag{5}$$

where g and h are the indexes of the number of onsets, λ_g^a, μ_h^a are the parameters, m is the number of onsets detected in the segment, and $\delta_{\{\cdot,\cdot\}}$ is Kronecker's delta. Due to the binary representation of onset, we can take account of intuitive features, such as whether the top of a segment is an onset, and the number of onsets in a segment can be detected.

2.3. Transition Functions

In our model, the duration and transition probabilities of Hidden Semi-Markov Models (HSMMs) are incorporated as the transition functions.

Let $d_n = t_n^e - t_n^s$ be the segment duration (s), r_n be the local tempo (s / beat), which is assumed to be constant in the segment, and let l_n be the chord length (beat) denoted in the score. The transition function is

$$\Psi(q_{n-1}, q_n) = \mathcal{N}(d_n; r_n l_n, \sigma^2)p_{s_{n-1},s_n}, \tag{6}$$

where $\mathcal{N}$ represents a Gaussian distribution with a mean of the expected duration $r_n l_n$ and variance of σ^2 , and p_{s_{n-1},s_n} are the HSMM transition probabilities.

Note that the tempo is assumed to be constant here for allowing inference using the Viterbi algorithm, but it is estimated adaptively during the alignment process and controls the transition functions dynamically.

3. LINEAR DYNAMICAL SYSTEM FOR TEMPO FLUCTUATION

To anticipate the future score position, we introduce a simple tempo model based on LDS, which is similar to an existing tempo model [9]. The tempo can fluctuate during a human performance, but, in general, it does not change considerably over a short period of time. Here, we assume that tempo can be constant locally. Thus, the tempo model is defined as

$$r_n = r_{n-1} + w_n, \quad (7)$$
$$d_n = r_n l_n + v_n, \quad (8)$$

where

$$w_n \sim \mathcal{N}(0, Q), v_n \sim \mathcal{N}(0, R), \quad (9)$$

for variance parameters Q and R. Equations (7) and (8) represent local tempo fluctuation and the observation process at inter-onset-intervals (IOI), respectively.

The simple linear model allows an efficient real-time decoding using a Kalman filter, which consists of prediction and correction steps, given the result of chord segmentation, which is the coupled sequence of IOI and the chord length.

4. DECODING ALGORITHM

4.1. On-line Approximation

Time-varying tempo can be estimated using a Kalman filter in an on-line manner. However, in our SCRFs, the most likely segmentation can be found using the Viterbi algorithm given the entire input audio. In score following, the input audio is given sequentially, thus we need some on-line approximations.

In [3–5], a greedy approximation that finds the most probable current score position is applied. However, it may cause estimation errors, particularly in polyphonic cases. To avoid this problem, we use a delayed-decision Viterbi algorithm for an on-line approximation that finds the most probable α-frame past-score position. Due to the utilization of future information, the algorithm can estimate a highly reliable score position. Although the idea is similar to [11], we use a backward strategy in the Viterbi algorithm and future anticipation from our explicit tempo model, as described in Section 3.

4.2. Delayed Decision and Future Anticipation

We now describe our future anticipation method for the score position. Let $\{\hat{s}_1, \ldots, \hat{s}_{t-\alpha}, \ldots, \hat{s}_t\}$ be the result of chord segmentation at frame t, $\{\hat{r}_1^{-1}, \ldots, \hat{r}_{t-\alpha}^{-1}, \ldots, \hat{r}_t^{-1}\}$ be the reciprocal of the tempo estimation (beats / s), and $\{b_1, \ldots, b_{t-\alpha}, \ldots, b_t\}$ be the sequence of score positions (beats) corresponding to the estimated chord sequence. The current or future score position is anticipated as

$$\hat{b}_t = b_{t-\alpha} + \int_{t-\alpha}^{t} \hat{r}_\tau^{-1} d\tau. \quad (10)$$

Here, if we assume that tempo is constant from $t-\alpha$ to t, the above equation can be approximated as

$$\hat{b}_t = b_{t-\alpha} + \hat{r}_{t-\alpha}^{-1}\alpha. \quad (11)$$

The delay time α is the important parameter in our decoding algorithm. However, the determination of this parameter is not straightforward. We test various values in our experiments.

Our score following algorithm based on this delayed decision and anticipation framework is summarized below. The algorithm is repeated for each time frame.

Step 1: Chord segmentation using a delayed-decision Viterbi algorithm for the input observation sequence

Step 2: Tempo estimation using a Kalman filter of the chord segmentation result

Step 3: Future anticipation using the results of Step 1 and Step 2.

5. EXPERIMENTS

5.1. Experimental Settings

We evaluate the robustness of our delayed decision and anticipation algorithm for polyphonic music signals using two datasets. The first contains 60 classical pieces with perfectly aligned MIDI data from the MAPS database [17]. These recordings are real-data played on a Disklavier piano (an acoustic piano equipped with MIDI input and output interfaces), which do not contain any tempo changes but are highly polyphonic. The second dataset consists of 50 pieces from the RWC Jazz database [18]. In contrast to MAPS, these recordings contain many tempo changes. Almost all of the recordings consist of multiple instruments including percussion. The ground truth is given by manually aligned MIDI files. For practical reasons, we only use these MIDI files due to the manual annotations might have a slight gap compared to the ground truth. To evaluate our algorithm correctly, we prepare perfectly aligned recordings by synthesizing these MIDI files with a YAMAHA XG WDM SoftSynthesizer, retaining all tempo changes. All recordings are re-sampled to 44.1 kHz monaural and analyzed with a 10 ms hop-size.

The algorithm is evaluated using three statistical measures: the Precision, Recall, and *F*-measure of the note onset recognition (in the same as in [15]). In these experiments, we report onsets if $\hat{b}_t$ reaches theoretical onset positions in the score. The onsets detected within a tolerance threshold corresponding to the reference onset time are accepted. The error tolerance is variously set to 100, 300 and 500 ms.

Model parameters are listed in Table 1. Note that we do not consider structural changes in music, such as skips of score events in these experiments. These can easily be adapted with a few modifications to the transition probabilities, as shown in [12], even using the on-line settings.

5.2. Results

Table 2 shows the results of onset detection for two databases with various delay times. In terms of the *F*-measure, the small delay time of 0.5 s obtained the highest result for the 100 ms tolerance threshold, showing an increase of over 30-% in both databases compared to no delay time. With a tolerance of 300

Table 2. Onset detection results (%) for 60 pieces from the MAPS database (top) and 50 pieces from the RWC Jazz database (bottom). The delay time is set to 0.0, 0.5, 1.0, and 1.5 s.

MAPS database

measure	Precision				Recall				F-measure			
α	0.0	0.5 s	1.0 s	1.5 s	0.0	0.5 s	1.0 s	1.5 s	0.0	0.5 s	1.0 s	1.5 s
100 ms	51.6	**78.9**	73.1	65.6	42.9	**78.7**	73.3	65.9	46.9	**78.8**	73.2	65.8
300 ms	89.0	91.8	**92.7**	92.3	73.1	91.5	**93.0**	92.7	80.3	91.7	**92.8**	92.5
500 ms	94.6	93.9	95.0	**95.3**	77.5	93.5	95.2	**95.6**	85.2	93.7	95.1	**95.4**

RWC Jazz data

measure	Precision				Recall				F-measure			
α	0.0	0.5 s	1.0 s	1.5 s	0.0	0.5 s	1.0 s	1.5 s	0.0	0.5 s	1.0 s	1.5 s
100 ms	37.2	**60.4**	54.7	47.6	25.4	**60.0**	54.8	48.1	30.2	**60.2**	54.7	47.8
300 ms	74.7	79.1	**80.3**	79.1	49.4	78.2	**80.3**	80.0	59.5	78.7	**80.3**	79.6
500 ms	85.0	84.0	85.8	**85.9**	55.4	83.1	85.9	**86.9**	67.0	83.5	85.8	**86.4**

Table 1. Model parameters for Segmental Conditional Random Fields and Linear Dynamical System.

SCRF parameters	
Chroma feature weight	$\lambda^c = 0.1$
Onset feature weight (1/3)	$\lambda_0^a = -0.3, \lambda_1^a = -0.01$
Onset feature weight (2/3)	$\mu_0^a = -0.3, \mu_1^a = -0.01$
Onset feature weight (3/3)	$\mu_h^a = -0.15h$ $(h \geq 2)$
Transition probabilities	$p_{i,j}$ = 1 only if $j = i + 1$ otherwise 0
Duration variance	$\sigma^2 = 0.18$ (s^2)
LDS parameters	
Tempo variance	$Q = 0.08$ (s^2/beat^2)
Inter-Onset-Interval variance	$R = 0.3$ (s^2)

ms, which is based on the Real-time Audio to Score alignment task in the Music Information Retrieval Evaluation eXchange (MIREX) contest [19], the results show an improvement of 11-% and 19-% in *F*-measure for the MAPS and RWC Jazz database, respectively.

The smaller the tolerance threshold, the greater is the delay time to obtain the results, which indicates that the delay time should be set according to the requirement of its application.

A large delay time (over 1.0 s) caused the results to worsen in the small tolerance of 100 ms. This situation arises from the trade-off between delayed decision and future anticipation accuracy. We might think that the large delay time, would enable higher accuracy in results, because of the availability of more future information about the input signal. However, the large delay time may cause anticipation errors. There are two reasons for this: the effect of tempo estimation errors, and the assumption that the tempo within the delay time is the same as the current tempo. The tempo estimation results are sometimes not reliable in these experiments. The accuracy is about 60-% with 4-% tolerance in both databases. Even if there are slight errors in the estimated tempo, the anticipation errors would increase in proportion to the delay time. However, it is worth mentioning that the results with small delay times obtained accurate results.

In the RWC Jazz database, our method obtained less accurate results than with the MAPS database. This is because the RWC recordings have more complexity than those in the MAPS, as mentioned in Section 5.1.

Using both databases, the Recall tends to be lower than the Precision with no delay time, particularly so for RWC Jazz. This can be explained by the fact that highly polyphonic music signals sometimes cause instabilities in the algorithm. However, the Recall is particularly improved using our delayed decision and anticipation algorithm. These results show the high robustness of our method for highly polyphonic music signals.

6. CONCLUSION

In this paper, we presented a robust on-line score following algorithm for polyphonic music signals based on a delayed decision and anticipation framework. The key features are our delayed-decision Viterbi algorithm, which finds highly reliable past positions utilizing future information, and that the future position can be anticipated using an adaptively estimated tempo thanks to our explicit tempo model.

Experimental results on polyphonic music databases showed significant improvements in alignment accuracy, even for highly polyphonic cases including tempo changes. It is worth mentioning that our delayed decision and anticipation framework can be used in existing dynamic programming-based score followers with an explicit tempo model. In future work, we intend to determine the delay time adaptively during a musical performance by considering the trade-off between the delayed decision and anticipation accuracy.

7. ACKNOWLEDGEMENT

This research was supported in part by the Ichihara International Scholarship Foundation.

8. REFERENCES

[1] R. B. Dannenberg and C. Raphael, "Music score alignment and computer accompaniment," *Communications of the ACM*, vol. 49, no. 8, pp. 38–43, 2006.

[2] A Cont, "Antescofo: Anticipatory synchronization and control of interactive parameters in computer music," in *Proc. of International Computer Music Conference (ICMC)*, 2008.

[3] P. Cano, A. Loscos, and J. Bonada, "Score-performance matching using hmms," in *Proc. of International Computer Music Conference (ICMC)*, 1999, pp. 441–444.

[4] N. Orio and F. Dechelle, "Score following using spectral analysis and hidden markov models," in *Proc. of International Computer Music Conference (ICMC)*, 2001, pp. 151–154.

[5] S. Dixon, "Live tracking of musical performances using on–line time warping," in *Proc. of International Conference on Digital Audio Effects (DAFx)*, 2005, pp. 92–97.

[6] T. Otsuka, K. Nakadai, T. Takahashi, T. Ogata, and H. G. Okuno, "Real-time audio-to-score alignment using particle filter for coplayer music robots," *EURASIP Journal of Advances in Signal Processing*, 2011.

[7] Z. Duan and B. Pardo, "A state space model for online polyphonic audio-score alignment," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 197–200.

[8] N. Montecchio and A. Cont, "A unified approach to real time audio-to-score and audio-to-audio alignment using sequential montecarlo inference techniques," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[9] C. Raphael, "Aligning music audio with symbolic scores using a hybrid graphical model," *Machine Learning Journal*, vol. 65, no. 2-3, pp. 389–409, 2006.

[10] A. Cont, "A coupled duration-focused architecture for realtime music to score alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 974–987, 2010.

[11] A. Arzt, G. Widmer, and S. Dixon, "Automatic page turning for musicians via real-time machine listening," in *Proc. of European Conference on Artificial Intelligence (ECAI)*, 2008, pp. 241–245.

[12] C. Joder, S. Essid, and G. Richard, "A conditional random field framework for robust and scalable audio-to-score matching," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 8, pp. 2385–2397, 2011.

[13] C. Joder, S. Essid, and G. Richard, "A conditional random field viewpoint of symbolic audio-to-score matching," in *Proc. of ACM Multimedia*, 2010, pp. 871–874.

[14] C. Joder, S. Essid, and G. Richard, "A comparative study of tonal acoustic features for a symbolic level music-to-score alignment," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 409–412.

[15] S. Dixon, "Onset detection revisited," in *Proc. of International Conference on Digital Audio Effects (DAFx)*, 2006, pp. 133–137.

[16] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 185–188.

[17] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[18] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical, and jazz music databases," in *Proc. of International Conference on Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.

[19] The International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL), "Real-time audio to score alignment (a.k.a score following)," `http://www.music-ir.org/mirex/wiki/Real-time_Audio_to_Score_Alignment_(a.k.a_Score_Following)`, Accessed 15 May. 2013.