

Real-time score following techniques

Jonathan Aceituno

1 Introduction

Computers have proven their relevance in a live musical performance context when used as helpers or spare performers. However, without a certain understanding of the musical context, the computer performer is barely more than a tape recorder. A crucial skill that has to be mastered for rich interaction between human and computer performers is score following. The principle of score following is to align a musical score to a live performance, so as to produce a sequence of matches between a timestamp and a score event. This way, the position in the score, and eventually the current tempo, is known in real-time during the performance and can be used for many purposes, like automatic accompaniment or automatic page turning.

The design of a real-time score following system throws down two main challenges. First, the stream produced from the live performance has to be correctly interpreted, in order to build a common representation for the score and the performance, so they can be paired together. Let us call this the interpretation problem. Second, these matches have to be found in real time and with a certain amount of flexibility and forgiveness. This is the alignment problem.

Other common problems are less often discussed in this research field, because they are related to a specific application. As far as automatic accompaniment is concerned, the system has to operate within harder real-time constraints for the accompaniment part to be played in time, leading to specific algorithms and optimizations. Tempo inference is also critical to these systems, thus the score follower can help to collect clues about the evolution of the tempo in the live performance.

Score following has been around for 27 years, it has known a tremendous evolution and a few major steps. The problem was introduced in 1984, simultaneously with two different approaches. The early

real-time score-performance alignment techniques dealt with a mostly monophonic MIDI performance input, but as voice-specific score followers were brought into consideration, the methods involved had to evaluate selected features extracted from a raw audio input. Stochastic frameworks brought more efficient audio-to-score alignment techniques, and only dynamic time warping based methods could compete with them. More recently, an anticipatory design has been proposed and stands for the most advanced proposal as for now.

This review will focus on the evolution of the methods that address the alignment problem in real time, although the two main problems are linked in such a way that every technique encompasses both of them. Every major technique will be detailed, and a particular attention will be drawn to the anticipatory design[3].

2 Early techniques

One of the two first approaches to score following defined the context of a synthetic performer whose tasks were to listen, perform and learn[19]. The listening part was related to the alignment problem and addressed it with a set of heuristics. Later on, a more complex alignment solution was designed, using a model of the performance inspired of the human brain, a dynamic programming alignment algorithm and a focus on learning through rehearsals[20].

The other first approach proposed an error-tolerant alignment technique with a bottom-up dynamic programming algorithm, where notes of the score are paired with notes of the performance, without any rhythmic consideration. In order to comply with real-time constraints, the algorithm would limit the calculations around a window[5]. More enhancements were done later. A pre-processing step was responsible for detecting

and grouping together incoming notes pertaining as a same complex event, like a chord or a trill. Ambiguous situations in the score following, like when the performer stops, were handled using several concurrent strategies, and when a strategy resolved the ambiguity, the others stopped.

A simpler technique was used at IRCAM in the early 90's, and consisted of sequential alignment trials between an incoming note and the next note on the score, with a preferred order and a skip list[15].

These techniques were using symbols like MIDI as performance inputs, but the need to apply score following to vocal performers led to the first audio-to-score alignment method[16]. As every audio-based score following system does, the audio input is divided into frames. The voice outputs variable pitches for a single note, mainly due to vibrato. However, an estimate of the instantaneous fundamental frequency is possible by doing a constant Q transform. After a note has been detected for a time long enough, a more precise fundamental frequency estimate is calculated. The precedent alignment system is used with the more precise estimate while the instantaneous frequency match is used to detect note onsets with a minimal delay.

3 Dynamic time warping

The principle of dynamic time warping is to match an input signal with a template signal without paying attention to rhythmic variations. The value of the input signal at time t is compared with values of the template signal around the last match in order to minimize the distance between the two values, for a given distance function. The aim is to find a continuous and monotonic path of matches that minimizes the global distance. This is done with a bottom-up dynamic programming algorithm. The template signals have to be generated depending on the nature of the input sequence.

The first alignment technique using dynamic time warping considered sequences of signal spectrums and several distances derived from the peak structure distance[11]. Memory complexity had been addressed by keeping only relevant paths. However, dynamic time warping is not a real-time algorithm and had to be adapted. A later technique proposed to compute only the minimal path to the current position on the score and to narrow calcu-

lations to a constant-sized window around the last match point, in order to perform online[6]. The sequences considered were vectors of positive spectral differences compared with an euclidean distance.

These simple techniques gave good results on polyphonic audio signals and do not require a training phase.

4 Stochastic models

The most popular stochastic approach is to model the score using hidden Markov models. A hidden Markov model is a stochastic process (X_1, \dots, X_n) where the probability of the presence in a state $q_t \in Q$ at time t does not depend on past states q_{t-2}, \dots, q_0 and can be written as $p(X_t = q_t | X_{t-1} = q_{t-1}, \dots, X_0 = q_0) = p(X_t = q_t | X_{t-1} = q_{t-1})$, where each state has an observation distribution, which represents the probability of generating a symbol $s \in S$. At any time, the state emits an observation according to the observation distribution, and then changes according to the state transition distribution, often represented as a finite state automaton. The evolution of the model is not observed and its parameters have to be discovered using the observations it emits. According to its transition matrix, a hidden Markov model presents a certain topology, which is usually left-to-right, because it naturally represents the way time goes on.

There are three classic problems to solve when using a hidden Markov model : evaluation, learning and inference[17]. The evaluation problem is to compute the probability of a certain observation sequence given a particular model. The learning problem is to adjust the model parameters in order to maximize the probability of observing a certain sequence. The inference problem is to find the most likely state sequence accounting for a given observation sequence and a given model, and it is solved using the Viterbi algorithm.

Score following using hidden Markov models was first motivated by the fact that a probabilistic model can cope well with audio features used to pair the score and the performance, that are often uncertain[1], at the cost of learning, ie. training the models with a database. The main concerns in using this approach for real-time score following are multiple.

Inference has to take place incrementally in real time, but the Viterbi algorithm needs future observations as well to perform. A common adaptation is to cut down the backtracking in order to determine the best path directly[1, 12]. Then, output observations have to be chosen carefully from audio features. A common approach is to combine spectral structure with other features like log-energy. In order to model these features from the score, a constant harmonic spectrum, an attack and a silence model are generated at each frame[12].

The most important concern, however, is that events in the score have to be modeled in a sequential way, but allowing for any particular state topology inside of events. The first approach was to consider three different score events : a note (with attack, sustain and rest states), a silence (with only rest states), and a *no-note* (anything that is noisy), with negative binomial implicit occupancy distributions accounting for durations[1]. Another approach was to use a two-level model, where low-level events were modeled with attack, sustain and rest states, but high-level events, corresponding to note entities, were modeled as two parallel states (a *n-state*, meaning the note has been played, and a *g-state*, meaning the note was missed)[12]. This way, different performer mistake patterns (wrong note, skip, extra note) can be handled for improved robustness. Further approaches only considered low-level events[18] and others used a combination of high-level events and events matching the beat structure of the score[8].

5 Anticipatory score following

Considering musical anticipation in computer systems has proven to be effective, not by modeling the anticipation phenomenon, but rather by using anticipation as a principle to build a system that depends on the past, the present and its expectations about the future[2]. An anticipatory alignment technique has been designed, as part of the Antescofo score follower, and has been reported to give outstanding results[3].

The intrinsic time structure of a few score events, like trills, can be very different of the time structure of the score. In order to accurately represent time in a hidden Markov model for simple events, like notes, or complex events, it is some-

times necessary to define a particular occupancy distribution for some states. However, state occupancy distributions in hidden Markov models are implicitly following an exponential density. Hidden semi-Markov models generalize the use of explicit occupancy distribution for states, which are semi-Markov, ie. depend not only on the former state, but also on the current occupancy distribution[10]. However, the forward-backward and the Viterbi algorithm have to be adapted and the huge complexity gain is very limiting for real-time applications. Hopefully, the number of actual semi-Markov states is low in most of the applications, and this hypothesis led to a hybrid Markov/semi-Markov model where any state can be either Markov or semi-Markov and where the main algorithms can be much more efficiently calculated[7]. Such a hybrid model is used to represent the score, and each score event is represented with one or more states with a particular topology.

The observation distribution for each state is modeled after frequency distribution templates made out of note information, such as pitch. Comparison between the current frame of the audio signal and the model is done by using the Kullback-Leibler divergence between the spectrum of the current frame as a frequency distribution and the observation distribution for a particular state.

As with other hidden Markov model alignment techniques, a cut down version of the Viterbi algorithm is used to infer the state sequence in real time. A notable exception with this technique is the use of an anticipative tempo agent to update occupancy distributions, compensating the lack of knowledge of the future states. The tempo agent uses an internal oscillator to model the current tempo and updates its parameters with the current position of the score follower. This way, the inference agent and the tempo agent interact with each other in a unique coupled fashion, accounting for an anticipatory mechanism. This technique, unlike some others, does not need an offline learning phase, as it can be done in real time.

6 Conclusion

Every alignment technique has its strengths and its flaws, but the earliest techniques cannot be compared with each other because they had different re-

quirements : one worked well with polyphonic symbolic inputs, whereas another one was only good at monophonic voice audio signals. Nevertheless, this distinction is now irrelevant, as every major technique is able to process at least polyphonic audio signals. The first attempt to design a standard evaluation protocol amongst score followers is very recent[4], and new score followers can now compete at MIREX.

To this point, as far as the author knows, only one new alignment technique has been proposed and consists of a switch between two strategies according to the reliability of the estimation of the current score position, given by a particle filter[13, 14]. However, further improvements to score following will be more likely to address other problems, such as the need for a rich and composer-friendly way to specify a score[9].

References

- [1] P Cano and A Loscos. . . . Score-performance matching using hmms. *In practice*, 1999.
- [2] A Cont. Modeling musical anticipation: From the time of music to the music of time. 2008.
- [3] A Cont. A coupled duration-focused architecture for realtime music to score alignment. *IEEE transactions on pattern analysis and machine intelligence*, 2009.
- [4] A Cont, D Schwarz, and N Schnell. . . . Evaluation of real-time audio-to-score alignment. *Proceedings of 8th . . .*, 2007.
- [5] RB Dannenberg. An on-line algorithm for real-time accompaniment. *Proceedings of the 1984 International Computer . . .*, Jan 1984.
- [6] S Dixon. An on-line time warping algorithm for tracking musical performances. *Proceedings of the International Joint Conference on . . .*, Jan 2005.
- [7] Y Guédon. Hidden hybrid markov/semi-markov chains. *Computational statistics & Data analysis*, Jan 2005.
- [8] A Jordanous and A Smail. Artificially intelligent accompaniment using hidden markov models to model musical structure. *MUSICAL STRUCTURE*, page 84, 2008.
- [9] S Lemouton and P Manoury. Suivi de partition, mise au point, perspectives. pages 1–9, May 2009.
- [10] KP Murphy. Hidden semi-markov models (hsmms). *unpublished notes*, 2002.
- [11] N Orio. . . . Alignment of monophonic and polyphonic music to a score. *Proceedings of the International Computer Music . . .*, Jan 2001.
- [12] N Orio. . . . Score following using spectral analysis and hidden markov models. *Proceedings of the ICMC*, Jan 2001.
- [13] T Otsuka, K Nakadai, T Takahashi, and T Ogata. . . . Two-level synchronization using particle filter for co-player music robots. *winnie.kuis.kyoto-u.ac.jp*, 2010.
- [14] Takuma Otsuka, Kazuhiro Nakadai, Toru Takahashi, Tetsuya Ogata, and Hiroshi G Okuno. Real-time audio-to-score alignment using particle filter for coplayer music robots. *EURASIP Journal on Advances in Signal Processing*, 2011:1–13, Jan 2011.
- [15] M Puckette. . . . Score following in practice. *Proceedings of the International Computer . . .*, Jan 1992.
- [16] M Puckette. Score following using the sung voice. *In Proceedings of the ICMC*, 1995.
- [17] LR Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [18] D Schwarz and A Cont. . . . Score following at ircam. *Proceedings of the 7th Inter*, Jan 2006.
- [19] B Vercoe. The synthetic performer in the context of live performance. pages 1–2, Sep 1984.
- [20] B Vercoe and M Puckette. Synthetic rehearsal: Training the synthetic performer. *Proceedings of ICMC*, Jan 1985.