A Unified Approach to Real Time Audio-to-Score and Audio-to-Audio Alignment Using Sequential Montecarlo Inference Techniques

Nicola Montecchio, Arshia Cont

A UNIFIED APPROACH TO REAL TIME AUDIO-TO-SCORE AND AUDIO-TO-AUDIO ALIGNMENT USING SEQUENTIAL MONTECARLO INFERENCE TECHNIQUES

Nicola Montecchio

University of Padova
Department of Information Engineering
Via Gradenigo, 6/B, 35131 Padova - Italy

Arshia Cont

Institut de Recherche et Coordination
Acoustique/Musique (IRCAM)
1, place Igor-Stravinsky, 75004 Paris - France

ABSTRACT

We present a methodology for the real time alignment of music signals using sequential Montecarlo inference techniques. The alignment problem is formulated as the state tracking of a dynamical system, and differs from traditional Hidden Markov Model - Dynamic Time Warping based systems in that the hidden state is continuous rather than discrete. The major contribution of this paper is addressing both problems of audio-to-score and audio-to-audio alignment within the same framework in a real time setting. Performances of the proposed methodology on both problems are then evaluated and discussed.

Index Terms— sequential montecarlo, particle filtering, real time systems, music alignment, score following

1. INTRODUCTION

A music alignment system aims at finding a mapping between two music signals, possibly of different nature, that associate all the parts of each signal to music events. In this paper, we focus our attention on the alignment of a streaming audio signal of a music performance, that is a signal fed as input to the system incrementally in real time, against a symbolic score (the digitized version of traditional music notation) or another audio recording.

Scientific research has focused over the years on real time alignment against a symbolic score (usually referred to as *score following*). Approaches have evolved from pattern matching techniques in the earliest works dating back to the mid-80s to Dynamic Time Warping (DTW) approaches and Hidden Markov Models (HMM); a review of these approaches, which deal mostly with monophonic signals, can be found in [1]. Similar techniques are also employed in other fields such as gesture following, where HMM map real time motion capture data to reference gestures [2]. Recent work on score following, focusing on polyphonic signals, also proposed hybrid graphical models [3] and Hidden Hybrid Markov/semi-Markov models [4] which deal with tempo (i.e., speed of performance) estimation explicitly, and represent the state of the art systems.

Audio to audio alignment received less attention, and has been investigated most notably in [5] which adopts a real time version of DTW. In particular, a common trend in many approaches to score following is to synthesize the reference score and to align the incoming audio against what is, effectively, an audio signal.

Our approach presents a unified methodology for the real time alignment of audio to both a symbolic score and an audio reference by exploiting sequential Montecarlo inference techniques, also known as particle filtering [6]. The advantages of our approach, besides the powerful statistical framework and inherent simplicity of the algorithm, are twofold: unlike most systems, tempo is an explicit parameter within the stochastic framework defined through musical motion equations; moreover, both symbolic and audio alignment problems can be formulated within the same framework by exploiting a continuous representation of the reference media as depicted in Fig. 1 for the symbolic case.

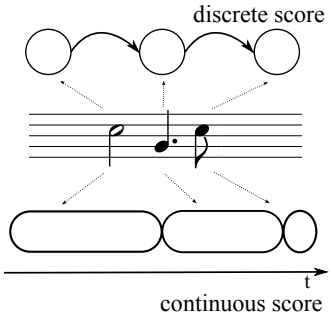


Fig. 1. From a discrete (one state per event) to a continuous (one region per event) score representation.

In the following sections we present in detail the methodology used for both score and audio alignment, and provide experimental evaluation on a hand-labeled test collection that is challenging for current state of the art software. We conclude with an overview of future research directions.

2. METHODOLOGY

Given a music stream and a reference medium, in the form of either a symbolic score or an audio recording, we formulate the alignment problem as a tracking problem, where the current position of the audio stream along the reference is modeled using traditional motion equations.

The system state is modeled as a two-dimensional vector $x = (s, t)$, representing the current position in the reference media and tempo respectively. In the case of audio to audio alignment, s is measured in seconds and t is the playback speed ratio, while in the case of a symbolic score reference, s is measured in musical time from the beginning of the score, and t in quarter notes per second (i.e. bpm/60). The incoming signal processing frontend is based on spectral features extracted from the FFT analysis of an overlapping, windowed signal representation, with hop size ΔT.

In order to use Sequential Montecarlo methods to estimate the hidden variable $x_k = (s_k, t_k)$ using observation z_k at time frame k, we assume that the state evolution is Markovian (i.e. depends only on the previous state and the current observation) and define the following quantities:

- $p(z_k|x_k)$ is the likelihood of observing an audio frame z_k given the current position along the reference s_k. We consider a simple spectral similarity measure, defined as the Kullback-Leibler divergence between the power spectrum at frame k and the power spectrum at time s_k in the reference audio, or, in the symbolic case, a template spectrum associated to the score event which is active at score position x_k. Past experience with realtime score following suggests that the power spectrum is a better feature for similarity computation than chroma, as it preserves octave information.

- $p(x_k|x_{k-1})$ is the state transition likelihood; we make use of tempo estimation in the previous frame and assume that tempo is circa equal:

$$p(x_k|x_{k-1}) = \mathcal{N}(\begin{bmatrix} s_k \\ t_k \end{bmatrix} \mid \mu, \Sigma)$$

$$\mu = \begin{bmatrix} s_{k-1} + \Delta T\, t_{k-1} \\ t_{k-1} \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_s^2\, \Delta T & 0 \\ 0 & \sigma_t^2\, \Delta T \end{bmatrix}.$$

  Intuitively, this corresponds to a performance where it is expected that the tempo is rather steady but can fluctuate; the parameters σ_t^2 and σ_s^2 control the variability of tempo and the possibility of local mismatches that do not affect the overall tempo estimate (e.g., a single note played with some delay).

- $q(x_k|x_{k-1}, z_k)$ is the particle sampling function. In our implementation this corresponds to the transition probability density function.

Sequential Montecarlo methods work by recursively approximating the current distribution of the system state using the technique of Sequential Importance Sampling: a random measure $\{x_k^i, w_k^i\}_{i=1}^{N_s}$ is used to characterize the posterior pdf with a set of points over the state domain and associated weights, and is updated at each time step as in Algorithm 1. An optional resampling step is used to address the *degeneracy* problem, common to particle filtering approaches; this is discussed in detail in [6, 7]. The decoding of position and tempo is carried out by computing the expected value of the resulting random measure as $\sum_{i=1}^{N_s} x_k^i w_k^i$.

Algorithm 1 SIS Particle Filter - Update step

for all $i = 1 \dots N_s$ **do**
  sample x_k^i according to $q(x_k^i|x_{k-1}^i, z_{k-1})$
  $\hat{w}_k^i \leftarrow w_{k-1}^i \frac{p(z_k|x_k^i)p(x_k^i|x_{k-1}^i)}{q(x_k^i|x_{k-1}^i, z_k)}$
  $w_k^i \leftarrow \frac{\hat{w}_k^i}{\sum_j \hat{w}_k^j}$
end for

3. EXPERIMENTAL RESULTS

The first set of experiments investigates the ability to adapt to tempo changes by using a synthesized note sequence, while the remaining ones make use of real music recordings, specifically a collection of Chopin's mazurkas annotated with the onset times for the events in the score; in particular, we used a subset of the collection created by C. Sapp for the Mazurka Project[1] consisting of 8 mazurkas for which 4 different performances were annotated. Due to space constraints we omit comparison with other approaches, relegating them to evaluation campaigns such as MIREX[2], and investigate instead specific situations that benefit from peculiarities of our approach.

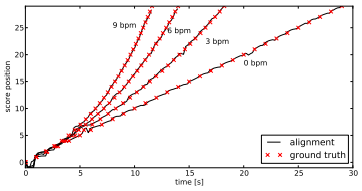
3.1. Synthetic Data Alignment Evaluation

A random monophonic note sequence was synthesized using a clarinet patch from an orchestral sample library. As the alignments plotted in Fig. 2 show, tempo increases linearly at every quarter note (Fig. 2(a)) or suddenly (Fig. 2(b)), starting from 60 bpm. We measured the average and maximum alignment error for the score events; such error is defined to be the delay or anticipation of the first detection of the event w.r.t. the nominal onset time. Results are summarized in Table 1.

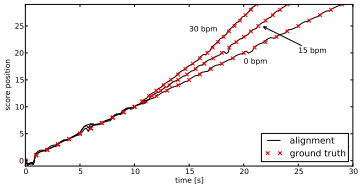
In all the tests a local misalignment can be seen around score position 5; this is due to the fact that two subsequent notes are repeated and no form of onset detection is used. Increasing the number of particles (to 2000) leads to a more robust tempo estimation and consequently avoids this problem. It should be noted that the average error is comparable to the analysis window hop size.

[1] http://www.mazurka.org.uk/

[2] http://www.music-ir.org/mirex/wiki/MIREX_HOME



(a) Incremental tempo increase



(b) Sudden tempo change

Fig. 2. Alignment to a synthetic track.

3.2. Audio to Score Alignment Evaluation

The symbolic scores were extracted from MIDI files downloaded from the Web, thus increasing the difficulty of the problem because of the many imperfections that are often present (i.e., skipped notes and incorrect notation); another significant issue is related to the intrinsic difficulty in formalizing music aspects like embellishments – which are extensively used in Chopin's music and realized differently by performers – in a format like MIDI.

Chopin mazurkas form a suitable test set for two reasons: they are complex polyphonic pieces challenging for state of the art score following softwares, and their executions are characterized by substantial tempo oscillations not explicitly notated in the score [8].

Results are provided in Table 2, using the same evaluation methodology and parameters that were used with the synthetic dataset; subsequently we treat particular cases in separate paragraphs for the mazurkas marked by an asterisk. A close analysis of the results, obtained inspecting the alignment plots, reveals that most of the times the maximum alignment error occurs on the very last beats of the pieces, characterized by a significant *rallentando*.

tempo change	avg. error (s)	max. error (s)
steady	0.12	0.71
3 bpm linear	0.13	0.63
6 bpm linear	0.09	0.34
9 bpm linear	0.09	0.23
15 bpm sudden	0.12	0.71
30 bpm sudden	0.11	0.52

Table 1. Audio to score alignment - synthetic data set.

	avg. error (s)	max. error (s)
Op. 6 n. 4	0.11	1.24
Op. 7 n. 2 *	0.16	2.65
Op. 17 n. 4	0.19	2.01
Op. 24 n. 2	0.19	3.89
Op. 30 n. 2	0.13	2.32
Op. 63 n. 3	0.18	2.89
Op. 67 n. 1	0.21	2.31
Op. 68 n. 3 *	0.34	4.42

Table 2. Audio to score alignment - Chopin's Mazurkas.

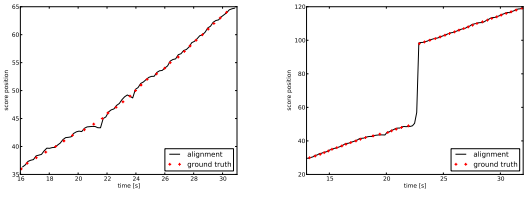
3.2.1. *Handling of Optional Repetitions*

Even though recorded interpretations of classical music usually respect the composer's instructions regarding repetitions, performers often choose to skip some of those repetitions, as is the case for one of the recordings of mazurka Op. 7 n. 2. The ability to follow optional repetitions can in principle be extended to the case where the performance can, at certain pre-defined points in the score, skip ahead (or go back) to other sections of the music, thus creating an *open form* interactive score, a common practice in contemporary music.

The particle filtering scheme is easily adapted to this situation, by allowing the particles to "jump" via a simple modification of the sampling step and transition pdf. Figure 3 compares two performances, where the skip of the repetition is clearly visible.

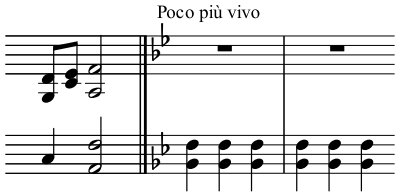
3.2.2. *Tempo Change Issues*

The analysis of the alignments of mazurka Op. 68 n. 3 pointed out a situation which is at the moment problematic for our score following system: between bars 31 and 32 there is a tempo change indicated by the composer, followed by 12 repetitions of the same chord that render an estimation of the new tempo not possible without some form of onset detection. As Figure 4 shows, the alignment is correct again when the melody starts, because of the added diversity to the harmonic content.

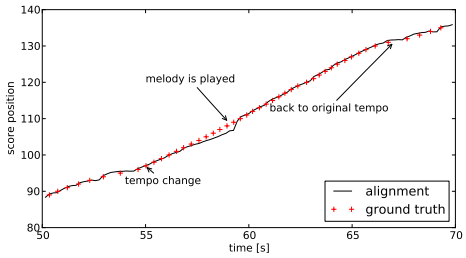


(a) Repetitions as in the score (b) Skipped repetitions

Fig. 3. Alignment of a performance with optional repetitions.



(a) score



(b) alignment

Fig. 4. Uncertainty in tempo estimation.

3.3. Audio to Audio Alignment Evaluation

For the audio to audio alignment experiments, the test collection of Section 3.2 was reused. In Table 3 we report the average and maximum alignment error for the score events; such error is defined to be the euclidean distance of the ground truth label from the closest alignment point. A similar parametrization as for the audio to score case was used, however only 200 particles are employed.

An interesting case is that of the repetitions in mazurka Op. 7 n. 2, investigated in Section 3.2.1 on the symbolic score following side. In the audio to audio case it is impossible to define possible jump points automatically, since the score is not known; however in some cases the alignment was nevertheless "correct", i.e., resembling the alignment curve of Figure 3(b); this is possible only if the variance of the transition pdf is not too low.

	avg. error (s)	max. error (s)
Op. 6 n. 4	0.18	0.59
Op. 7 n. 2 *	0.18	1.06
Op. 17 n. 4	0.17	2.14
Op. 24 n. 2	0.29	9.41
Op. 30 n. 2	0.13	0.58
Op. 63 n. 3	0.15	1.23
Op. 67 n. 1	0.16	1.36
Op. 68 n. 3	0.19	2.12

Table 3. Audio to audio alignment - Chopin's Mazurkas.

4. CONCLUSION AND FUTURE WORK

A system for the alignment of a music audio stream to both symbolic and audio references using a unified methodology was presented and its validity demonstrated by its application to a collection of polyphonic music performances.

As anticipated in the analysis of experimental data, the subject of future research will revolve around the ability to follow open-form interactive scores, a priority that will have immediate application because of the already active collaboration with composers on works that are going to exploit this aspect explicitly in their artistic conception.

Another important aspect, currently under investigation, is the dynamic tuning of the model parameters σ_s, σ_t using extended Kalman filtering techniques to increase accuracy.

5. REFERENCES

[1] N. Orio, S. Lemouton, and D. Schwarz, "Score following: state of the art and new developments," in *Proceedings of the conference on New Interfaces for Musical Expression*, Singapore, 2003, pp. 36–41.

[2] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. H. Rasamimanana, "Continuous realtime gesture following and recognition," in *Gesture in Embodied Communication and Human-Computer Interaction*. 2009, vol. 5934 of *Lecture Notes in Computer Science*, pp. 73–84, Springer.

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

[4] A. Cont, "A coupled duration-focused architecture for real-time music-to-score alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 974–987, 2010.

[5] S. Dixon and G. Widmer, "Match: A music alignment tool chest.," in *Proceedings of the International Conference on Music Information Retrieval*, 2005, pp. 492–497.

[6] M. S. Arulampalam, S. Maskell, and N. Gordon, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, pp. 174–188, 2002.

[7] R. Douc, O. Cappe, and E. Moulines, "Comparison of resampling schemes for particle filtering," in *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2005, pp. 64–69.

[8] P. Grosche, M. Müller, and C. S. Sapp, "What makes beat tracking difficult? a case study on Chopin's mazurkas," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 2010.