AN AUDIO TO SCORE ALIGNMENT METHOD USING VELOCITY-DRIVEN DTW FOR MIREX 2014

O. Romaní Picas, J.J. Carabias-Orti
Music Technology Group, Universitat Pompeu Fabra, Barcelona

ABSTRACT

The problem of audio to score alignment has been addressed from the beginning of the 80's and is nowadays well understood, leading to high accuracies even for complex polyphonic musical inputs. Traditionally, the evaluation metrics rely on the distance between the ground truth and the estimated note onsets, considering a fixed tolerance threshold (e.g. 200 ms). This criterion is suitable for many applications such as page turning or informed sound source separation. However, other applications that need the synthesis of the output, as automatic musical accompaniment systems, require a more advanced alignment. The aim of the algorithm presented in this paper is to guide the alignment process from a more musical perspective in order to provide an output that could be used in an automatic musical accompaniment system.

The work comes from the Sound and Music Computing master thesis offered by the Music Technology Group and is entitled "A novel audio-to-score alignment method using tempo-driven DTW".

1. SYSTEM DESCRIPTION

A score follower is an algorithm that estimates a score position of an input audio in an online fashion. Usually, the score is given in a MIDI file and the input audio in a WAV file. The proposed system has two different stages; preprocessing and alignment. The system block diagram is shown in Figure 1.

1.1 Preprocessing stage

The fist stage (preprocessing) is based in the work presented in [1]. First of all, the MIDI is analyzed to identify the combinations of concurrent notes in the score. Then we define the "states" as the transitions between these combinations. Then, the MIDI is synthesized using Timidity++ and the FluidR3 GM soundfont. From this audio, the spectral pattern for each unique combination of notes is learned using a supervised Non-Negative Matrix Factorization (NMF) method.

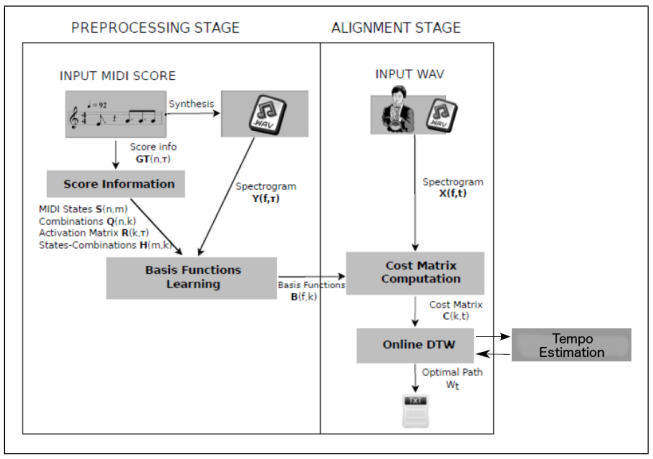

Figure 1. Proposed modification of the alignment stage presented in [1]

1.2 Alignment stage

In the second stage (alignment), the patterns learned for each state are projected onto the "real" audio using the method proposed in [2]. This process leads to a distortion matrix D (i.e. a cost matrix) where the alignment is computed. Finally, the minimum path is chosen using an online DTW method that accounts for tempo variations. The proposed DTW is based on the works form [3] and [4].

It should be noted that since the algorithm is online, the decisions are made using only the forward stage. Concretely, the minimum cost path $W = w_1, ..., w_k, ..., w_K$ is obtained by choosing the minimum cost path as follows:

$$W_k = \min\{C(i_k, j_k)\} \tag{1}$$

where C is the warping matrix $\mathbf{C}$ is filled recursively as:

$$C(i,j) = \min \left\{ \begin{array}{c} C(i, j-c_j) + \sigma(0, c_j) D(i,j) \\ C(i-c_i, j) + \sigma(c_i, 0) D(i,j) \\ C(i-c_i, j-c_j) + \sigma(c_i, c_j) D(i,j) \end{array} \right\} \tag{2}$$

where c_i and c_j are step size at each dimension and range from 1 to α_i and 1 to α_j, respectively. α_i and α_j are the maximum step size at each dimension. Parameter σ controls the bias toward a concrete direction. $C(i,j)$ is the cost of the minimum cost path from $(1,1)$ to (i,j), and $C(1,1) = D(1,1)$.

In this work, we aim to restrict the path according to the tempo performance of the real audio by using two different

type of restrictions to the minimum path search; *Global restrictions* and *Tempo-driven constraints*.

On one hand, the global restrictions control that the duration of a state not vary more than $1/4$ and 4 times the duration indicated in the score. Therefore, we define $c_i = c_j = 4$ and we restrict the search in three different ways for each state.

1. **time is below** $stateduration/4$: the search is restricted to the duration of the state
2. **time is between** $stateduration/4$ **and** $stateduration*4$: the search is restricted to the current and next state
3. **time is above** $stateduration * 4$: the search is restricted to the next state.

where the state duration is defined as the number of frames of the state in the original MIDI score.

On the other hand, the tempo-driven constraints impose some bias in the path construction by penalizing those steps that lead to tempo values different than the one analyzed during a previous period of time.

First of all, we have to decide the beginning and the end of the different time periods at which we are going to analyze the tempo. To do so, we compute the correlation between the spectral patterns of the consecutive states and then we look for the minimum peak values at the correlation vector. Those states with minimum correlation values are defined here as anchor points.

Finally, the tempo is computed between two anchor points as the mean value of the steps (c_i, c_j) chosen for the minimum cost path ($W(i_k, j_k)$) and the bias control parameter $\alpha(c_i, c_j)$ is updated accordingly.

2. RESULTS

Please go to http://www.music-ir.org/mirex/wiki/2014:Real-time_Audio_to_Score_Alignment_ (a.k.a._Score_Following) _Results#Summary_Results

3. REFERENCES

[1] J. Carabias, F. Rodriguez, P.Vera, P. Cabañas, F.J. Cañadas and N.Ruiz : "MIREX 2013 Real Time Audio to Score Alignment: A real-time nmf-based score-follower" Music Information Retrieval Evaluation eXchange, 2013.

[2] J.J. Carabias-Orti, F.J. Rodriguez-Serrano, P. Vera-Candeas, F.J. Caadas-Quesada, N. Ruiz-Reyes," Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription", *Engineering Applications of Artificial Intelligence*, Volume 26, Issue 7, August 2013,

[3] R. Turetsky and D. Ellis (2003) Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses *4th International Symposium on Music Information Retrieval ISMIR-03*, pp. 135-141, Baltimore, October 2003.

[4] S. Dixon, An On-Line Time Warping Algorithm for Tracking Musical Performances, *Proceedings of the International Joint Conference on Artificial Intelligence*, Edinburgh, August 2005, pp 1727-1728.