REAL TIME AUDIO TO SCORE ALIGNMENT BASED ON NLS MULTIPITCH ESTIMATION (1) MIREX 2010

F.J.Rodriguez-Serrano, P.Vera-Candeas,J.J.Carabias-Orti, P.Cabañas-Molero, N.Ruiz-Reyes
Telecommunication Engineering Department, University of Jaen
Polytechnic School, Linares, Jaen, Spain
{fjrodrig,pvera,carabias,pcabanas,nicolas}@ujaen.es

ABSTRACT

Real-time audio to score aligment is a valuable application in some music areas. This goal has been addressed from different point of views during the last years. Multipitch estimation was one of the first processing algortihm used for audio to score alignment. Here, a multipitch estimator is utilized to inform about the probability that a set of notes is active in a signal frame. In this way, the system is able to associate the states of score to the time of the performance. This approach is designed to be low complexity in order to be used in real-time applications.

1. INTRODUCTION

The alignment between the performance and the score is usually addressed in a two step algortihm. First audio and score are analyzed to extract the probability that a certain audio frame belong a score state. Second, these probabilities are utilized to perform the correct link between audio time and score states.

The analysis between audio and score has been addressed with well-known tools from audio signal processing. In [1] [2], frequency techniques based on Short-Time Fourie Transform are used. A correlation between each audio frame transform and the ideal harmonically-related frequency for each given set of notes from the score is computed [2]. In [3] an onset detection followed by pitch detection is proposed. In [4], note onsets are computed by decomposing the audio signal into spectral bands corresponding to the fundamental pitches and harmonics, followed by computation of the positions of significant energy increases for each band.

In order to perform a link between audio time and score states two approaches are generally utilized: Dinamic Time Warping (DTW) [5] or probabilistic models based on Hidden Markov Models (HMMs). DTW has predominantly been used for offline techniques being of low complexity and giving promising results in different scenarios [2] [6]. Typically HMMs and other graphical have been used for online applications [7] [8] [9] [10].

In this proposal, a NLS multipitch estimator is used as a processing tool to obtain probabilities between audio time and score states. Then, a sub-optimal approach of DTW has been designed in order to take decisions with limited latency in order to allow real-time decision to the score follower.

2. MULTIPITCH ESTIMATION

One of the simplest multipitch estimators is nonlinear least-squares method (NLS). The NLS estimates are obtained as the set of fundamental frequencies that minimizes the 2-norm of the difference between the observed signal and the signal model [11]. The signal model includes a set of harmonically-related frequencies at integer multiples of the each fundamental frequency ω_k. Due to this property, NLS method can be implemented efficiently for a linear grid search over a set of ω_k using a fast Fourier transform (FFT) [12].

One of the main problems of NLS [12] is the estimation of the order or the number of partials that belong to each fundamental frequency. Some solutions have been proposed in the literature for order estimation in multipitch methods [13]. Here, a solution based on the perceptual significance of each partial is proposed. The perceptual significance of frecuency peaks is computed following the perceptual model of [14]. This perceptual models gives a value larger than one to those peaks audible and lower than one to those peaks below the hearing threshold. A cumulative product of the perceptual significance of the partials belonging to the each evaluated fundamental frequency is calculated. The order for each fundamental frequency is suppossed to be equal to the partial position at which the cumulative product presents the maximum. With the idea of evaluating, the benefits of proposed order estimation two different approaches are presented at the competition: with and without the order estimation. The version without order estimation supposes a fixed order of 9 partials for all fundamental frequencies.

Finally, the algorithm to obtain the probability that a score state (a set of notes) is played at a audio frame is the following:

- Obtain the FFT of the frame.
- Compute the amplitudes for all spectral peaks of the frame.
- Compute the perceptual significance for all spectral peaks of the frame.
- For each note in the current score state, estimate the order (and therefore the number of active partials).
- Sum all the active partials belonging to the set of notes to obtain the salience of the current score state.
- Compute the probability as the division of the salience of the current score state and the total salience of the frame (computed as the sum of all spectral peaks of the frame).

This computation must be done for each audio frame and for each state of the score, but the low complexity of NLS estimator and the used perceptual model allows to obtain the probabilities in a real-time execution.

In this way, a similarity matrix S that contains the probabilities for each state of the score (columns) and audio frame (rows) is given to the next block of the system in order to obtainb a relation between score and performance in real time.

3. DYNAMIC TIME WARPING

In this section the similarity matrix post-processing is described. Dynamic Time Warping (DTW) algorithm is used in order to find the matching between MIDI frames and temporal frames. The similarity matrix is compose as described in section 2, and it is the input for the DWT block.

DTW is a kind of dinamic programming (DP) [15]. DP consists of two stages, forward and traceback. In the forward step, it calculates the lowest-cost path to all of the points neighbors plus the cost to get from the neighbor to the point - in this case, $S_{max} - S(i, j)$, where S_{max} is the largest value in the similarity matrix. This makes the cost for the most similar frame pairs be zero, and all other frame pairs have larger, positive costs. The object of this stage is to compute the cost of the best path to point $(N1, M1)$ recursively by searching across all allowable predecessors to each point and accumulating cost from there. In the traceback stage, we find the actual path itself by recursively looking up the point providing the best antecedent for each point on the path.

Unconstrained DTW shold allow the path to go horizontally and vertically [16], but for this application, this may be constrained. Horizontal pathes would be a non planned (at the midi file) stop at the performance. This is possible but not very frequent. Vertical pathes would be a infinite playing velocity, this is imposible so it should be constrained. Then, a change matrix is given to the DTW where a vertical or horizontal path is penalized with a hihg cost, while the diagonal has a low one.

In order to have a real time system, it is not possible to wait until the end of the file to select the best path along the similarity matrix. Then a variation of it has been implemented. It takes sub-matrixes along real time and it applies the DWT algorithm to them. These are increasing sub-matrixes that contain information from the beggining time to the current time position.

The sub-matrixes increase is of 1 second along time frames and the same along midi frames. This increasing is added from the last estimated samples at the previous DWT execution, so matrixes grow up folowing the estimated sub-path.

At each iteration a parcial path solution is achieved, from the origin to the actual time. However only a strech is taken and added to the complete solution, this is from the last decision time up to 20% before now, because there is an information limitation beacause of the end of the matrix for the last samples, this will be solved at the next step where these samples will be estimated better.

Then, the complete solution of the path along the similarity matrix is composed with each strech extracted from each sub-path calculated at every sub-matrix. This path is the estimation of which temporal frame match with which midi frame. At each second, the new stecht is available at the output. A minimun latency is present because of the 20% of samples that are not taked into account from the new sub-path, because of it latency, can be from $200ms$ to $1000ms$.

The new output times are obtained by interpolating original times over the estimated path line.

4. REFERENCES

[1] Soulez, F.; Rodet, X. and Schwarz, D., "Improving Polyphonic and Poly-Instrumental Music to Score Alignment," *4th International Conference on Music Information Retrieval*,ISMIR 2003, 2003, 143-148.

[2] Turetsky, R. and Ellis, D., "Ground-truth transcriptions of real music from force-aligned midi syntheses," *4th International Conference on Music Information Retrieval*,ISMIR 2003, 2003.

[3] Arifi, V.; Clausen, M.; Kurth, F. and Muller, M., "Automatic Synchronization of Music Data in Score-, MIDI- and PCM- Format," *4th International Conference on Music Information Retrieval*,ISMIR 2003, 2003.

[4] Muller, M.; Kurth, F. and Roder, T., "Towards an Efficient Algorithm for Automatic Score-to-Audio Synchronization," *5th International Conference on Music Information Retrieval*,ISMIR 2004, 2004.

[5] Rabiner, L.R. and Juang, B.H., "Fundamentals of Speech Recognition," *Prentice*, 1993.

[6] Dannenberg, R.B. and Hu, N., "Polyphonic Audio Matching for Score Following and Intelligent Audio Editors," *Proceedings of the 2003 International Computer Music Conference*, 2003.

[7] P. Cano, A. Loscos, and J. Bonada, "Score-performance matching using HMMs," *Proceedings*

of the International Computer Music Conference, 1999,pp 441-4.

[8] N. Orio and F. Dchelle, “Score following using spectral analysis and hidden Markov models,” *Proceedings of the International Computer Music Conference*, 2001,pp 151-4.

[9] C. Raphael, “A hybrid graphical model for aligning polyphonic audio with musical scores,” *Proceedings of the International Computer Music Conference*, 2004,pp 387-94.

[10] P. Peeling, T. Cemgil, and S. Godsill, “A probabilistic framework for matching music representations,” *Proceedings of the International Computer Music Conference*, 2007,pp 267-72.

[11] S. M. Kay, “Fundamentals of Statistical Signal Processing: Estimation Theory,” *Prentice-Hall*, 1993.

[12] M. G. Christensen, P. Stoica, A. Jakobsson and S. H. Jensen, “Multi-Pitch Estimation,” *Signal Processing*, vol. 88(4), pp. 972-983, 2008

[13] M. G. Christensen and S. H. Jensen, “Variable order harmonic sinusoidal parameter estimation for speech and audio signals,” *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2006.

[14] S. van de Par, “A perceptual model for sinusoidal audio coding based on spectral integration,” *EURASIP Journal on Applied Signal Processing*, v.2005 n.1, p.1292-1304, 1 January 2005

[15] Gold, B. and Morgan, N., “Processing and Perception of Speech and Music,” *Speech and Audio Signal Processing:*, John Wiley and Sons, Inc., New York,1999.

[16] Song, Xiang, Chu, Chengbin and Nie,, “A heuristic dynamic-programming algorithm for 2D unconstrained guillotine cutting,” *Proceedings of the IADIS International Conference on Applied Computing*,Lisbon, Portugal, 23-26 March 2004.