POLYPHONIC SCORE FOLLOWING USING ON-LINE TIME WARPING

Robert Macrae
Centre For Digital Music
robert.macrae@elec.qmul.ac.uk

Simon Dixon
Centre For Digital Music
simon.dixon@elec.qmul.ac.uk

ABSTRACT

This paper describes the method used in MidiMatch, an On-Line Time Warping approach to Score Following, that was submitted to the Mirex 08 Score Following task.

1 INTRODUCTION

Score Following refers to the synchronisation of audio and musical scores in real-time for purposes such as page-turning or automatic accompaniment. Dynamic Time Warping is a method for finding the best path through a matrix of costs. DTW has become common for off-line synchronisation methods as in [6, 10, 11], however, due to quadratic time and memory costs, it is seen as unsuitable for real-time applications. On-Line Time Warping and the Match application by Dixon [4, 3] have made DTW in real-time possible by iteratively matching segments of a similarity matrix when they are received, resulting in linear time and memory costs. While Match synchronises two pieces of audio using OLTW based on derivative spectral features, MidiMatch has extended this application to work with MIDI and audio files to produce a list of linked score and audio events.

Off-line DTW score synchronisation methods, that are based on spectrum/chroma features, will synthesise the musical scores, generating the expected audio and then taking the spectral features from this audio. As this delay is unsuitable for real-time applications, MidiMatch generates spectral features from the MIDI events using pre-computed mappings trained on various synthesised MIDI notes.

2 DTW VS HMMS

Although Dynamic Programming [2], Bayesian Networks [7] and Graphical Modelling [8] have all been used, Hidden Markov Models stands out as the current de-facto method for Score Following [1]. HMMs probabilistic nature matches Score Following well, as the performance of a piece is rarely an exact replication of the score. Computationally more efficient than standard DTW, HMMs require training and can be less suited to polyphonic music. For a more in depth comparison of DTW and HMMs see Durbin et al [5].

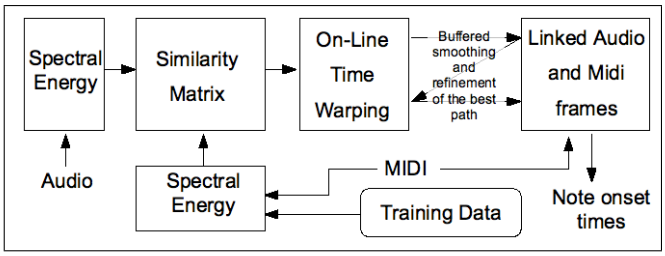

Figure 1. An overview of the methodology.

3 METHOD

MidiMatch takes the spectral information of audio by using a short time fourier transform of 20 ms audio frames and then maps these to a non-linear frequency scale. Note onsets are emphasised by discarding all but the positive spectral difference vectors and near silent frames are discarded. The MIDI events are mapped to spectral data in the same frequency scale as the audio and are then further adjusted to resemble real audio. The OLTW iterates through the available spectral data by using an euclidean distance metric of each pair of frames spectral difference. This best-path is then refined and smoothed using first a backward pass and then another forward pass bound by a Sakoe-Chiba [9] like path constraint based on the initial OLTW. The segment undergoing the repeated DTW acts as a buffer between the audio received and the aligned data. When two matching frames have been finalised and one of the frames corresponds to an MIDI event, the audio frame of the note onset is known. Figure 1 shows an overview of the the processes involved.

4 MAPPING MIDI EVENTS TO SPECTRAL FEATURES

The spectral models for the MIDI events are trained on pre-computed frequency values based on synthesised MIDI notes. For every instrument in the standard MIDI program list and every MIDI note from 0 to 127, the distribution of the spectrum is stored. Using these distributions and the velocity of the MIDI events, the mapped spectral information closely resembles similar audio.

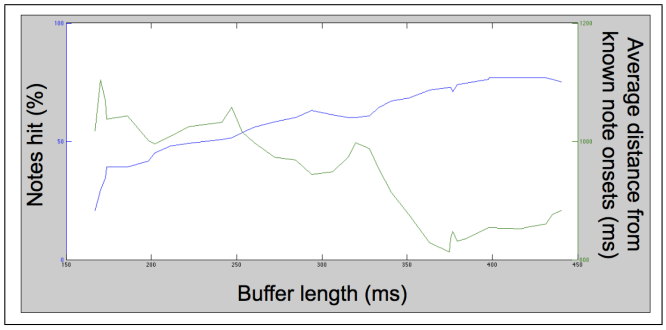

Figure 2. The trade-off between latency and accuracy.

5 OPTIMISATIONS

Adjustments had to be made to the derivative spectral features of the audio data to allow for a closer similarity to that generated by the MIDI. Parameters that affect the distortion caused by frames of background noise were evaluated and the MIDI spectral data was weighted to better simulate real audio. Rather than taking the first matching audio frame, a small latency was introduced to allow the backward and smoothing paths to refine the linked frames and this created a trade off between delay and accuracy as seen in Figure 2.

6 EVALUATION

Acquiring ground truth data for testing synchronisation methods is difficult and often requires hand labelling audio onset events. One set of data available was the reverse conducted mazurka data from the Centre for the History and Analysis of Recorded Music (`http://www.mazurka.org.uk/`). This data set consists of multiple recordings of monophonic piano pieces and lists of note onset times. Also used was a small set of samples from the Mirex task (used to chart the trade off between accuracy and buffer in Figure 2).

7 CONCLUSIONS AND FUTURE WORK

MidiMatch represents an early implementation of the OLTW method for use as a score tracker. Testing is still in the initial stages and there is clearly room for improvement in accuracy. The application may not be particularly suited to the Mirex test data as the samples suggest this is mainly monophonic and the MIDI files do not specify any instruments. This would suggest more generalised chroma features would work better and onset detection algorithms would more easily identify when an event has occurred than they would in polyphonic data. Apart from adjusting the spectral information to favour onsets using only positive differences, MidiMatch does not implement an onset detection method. This is now required to improve the accuracy of matched audio frames beyond the resolution of the window size.

The MIDI mapping data was generated from synthesised MIDI, however there are still audible differences in synthesised MIDI and recorded audio clips. This suggests using real audio may improve the mapping data. Other optimisations in efficiency are needed as the current latency is not ideal for a real-time application. This will be restricted by the buffer, which is required for accuracy.

Despite the delay, the overall application synchronises music of 20 seconds length in less than 3 seconds when run offline. This suggests it would be possible to run multiple matches concurrently to follow divergent possible paths, use multiple resolutions or combine with source separation to split polyphonic music into separate matches of single instruments. We hope to present an improved version of MidiMatch at Mirex 09 based on these improvements.

8 REFERENCES

[1] A. Cont and D. Schwarz. Score following at ircam. In *Mirex Submission*, 2006.

[2] R. B. Dannenberg. An on-line algorithm for real-time accompaniment. In *International Computer Music Conference*, pages 193–198, 1984.

[3] S. Dixon. Live tracking of musical performances using on-line time warping. In *Proceedings of the 8th International Conference on Digital Audio Effects*, pages 92–97, Madrid, Spain, 2005.

[4] S. Dixon and C. Widmer. Match: A music alignment tool chest. page 6, 2005.

[5] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. 1998.

[6] M. Muller and D. Appelt. Path-constrained partial music synchronization. In *ICASSP*, 2008.

[7] C. Raphael. A bayesian network for real-time musical accompaniment. page 7, 2001.

[8] C. Raphael. A hybrid graphical model for aligning polyphonic audio with musical scores. 2004.

[9] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), 1978.

[10] S. Salvador and P. Chan. FastDTW: Toward accurate dynamic time warping in linear time and space. In *Workshop on Mining Temporal and Sequential Data*, page 11, 2004.

[11] F. Soulez, X. Rodet, and D. Schwarz. Improving polyphonic and poly-instrumental music to score alignment. In *ISMIR*, pages 143–148, 2003.