

MATCH: A Music Alignment Tool Chest

By Simon Dixon and Gerhard Widmer

This paper looks at a piece of algorithm that allows for the alignment of two separately performed recordings of the same musical piece. It is robust and can tolerate moderate variations in tempo and rubato between the two recordings. The effort builds off of earlier work to align speech segments. The two main parts of the algorithm are the dynamic time warping (DTW) which makes use of an artificial cost function for comparing similarity between spectra of different parts of each recording.

In order to make processing easier, the two files are dividing into 20 ms frames. The algorithm will later compare the files and align them frame to frame. In order to compare the similarity between frames of each recording, a windowed Fast Fourier Transform is used to represent the spectral information contained within each frame. The 2048 point transform used is too high in dimension for practical processing and thus the frequencies are transformed to a linear-log scale which also imitates the sensitivity of the human ear. To do this, the lowest frequency bins (up to 370 Hz) are linearly mapped to the new scale while the bins of higher frequency are summed together into a logarithmic semitone scale upto 12.5 KHz. The resulting vector of spectral information for each frame is thus reduced from 2048 to 84. Next, the difference between adjacent "FFT vectors" is taken to form a derivative like vector with the negative components truncated. The final vectors represent the positive increase in spectral energy between successive frames. Finally, when comparing frames between recordings, the simple Euclidean distance is taken between the corresponding difference-vectors (identical frames would have zero distance).

The distance calculated between frames is used as a cost function representing the penalty for aligning the two frames i and j . The DTW algorithm is a dynamic programming technique which starts at the end of each recording and works backwards towards the beginning of each recording calculating the potential cost of aligning each pair of frames. At each step, the cost of moving backwards towards the start in each of the three possible directions is evaluated. The move with the lowest cost is chosen until the beginning is reached. The two dimensional path taken through both files represents the alignment between files that results from the least total cost.

If the algorithm is allowed to explore the entire alignment space, the algorithm would quadratic in time with the length of the sum of each file. An improvement in efficiency was made by restricting the search space to be within a fixed distance of the diagonal. This resulted in linear time efficiency.

The authors tested the algorithm on 3 set of data: a precision test using a computer-monitored piano capable of extracting beat onset times, CD recordings with annotation automatically marked with Beatroot and CD recordings with manually annotated beat timing. The average alignment error among the first (and most precise) test was 23 ms, well below the threshold of human hearing. For the BeatRoot-annotated recordings the average error was 64 ms.

The tool proves useful for automatically aligning different performances of the same music. The test data showed that it was fairly robust and could be used on different styles of music. The results are impressive considering that only low-level audio information was used to align different parts. Using this algorithm with the help of heuristics from a higher-level of audio information (such as beat and measure information) might improve the accuracy of such an alignment system.