

Review of “MATCH: A Music Alignment Tool Chest”

Abhijit Bhattacharjee

SIMON DIXON AND GERHARD WIDMER PRESENT MATCH, a set of software tools for aligning disparate recordings of a specific piece of music. A tool such as this can be invaluable for students of music who wish to compare different renditions of a musical work, because they must otherwise manually search for, and tag, similar instances of musical themes even in bodies of random access digital media such as CDs or MP3 files. This is because indexing provided by the producer is typically limited to the track level (or beginning of the piece), whereas musicology researchers and others may require a finer grain of indexing ability in order to contrast how two different cellists perform a particular phrase in a piece, for example.

MATCH uses a dynamic time warping (DTW) algorithm to align two different performances. Traditionally, DTW algorithms have quadratic complexity, which is challenging to compute as data sequences become longer. Therefore, the authors first innovated upon this process to convert it to a linear complexity algorithm, which is far easier for computers to handle and consumes relatively little calculation time even for large data sequences.

In a DTW algorithm, two “paths” of music are aligned by comparing pairs of points along each path to check for similarity. Typically, any two points along either path is available for comparison. However, the authors here restrict the paths to search only pairs of points that are close to each other—only one point interval away in any direction (horizontal, vertical, or diagonal). In conjunction with limiting the total search width, this reduces the algorithm’s complexity to linear time.

Once such a path is calculated, the audio file (which is divided into intervals called *frames*) is analyzed frame by frame for alignment. The main criterion for matching two audio frames is their spectral energy. A custom FFT scale is used for this computation. At the lower end, frequencies are grouped and separated linearly—e.g. 50 Hz, 60 Hz, 70 Hz, etc. Above 370 Hz, frequency groups are divided logarithmically. The authors claim that this method is more representative of how the human ear perceives sound. After such division, energy between each frequency group is summed and accumulated into bins, which are then used as features in a rudimentary, nearest-distance-based classifier.

The end product of all this computation is that for every time instance in the reference audio file, there is a corresponding “matched” time instance in each comparison file. Thus, events in the musical timeline of the reference piece can easily point to their counterparts in a comparison piece. In Dixon and Widmer’s testing, successful matching was performed in 681 out of 683 test cases, with average alignment error of 41 milliseconds. They postulate that better performance would be obtained with higher level features than spectral energy; features such as actual pitch could achieve even better alignment results. Unfortunately, the extraction of such features is not reliable enough in polyphonic music, therefore, further investigation will be required to handle high level attributes of music. Additionally, the MATCH system hiccups in cases where one performer repeats a section or phrase, where another does not; this is because the search bandwidth is only 5-10 seconds long. This would require a complete search amidst all sections of the files, rather than a windowed search strategy, making the process much more inefficient computationally.