

Jonathan Mooser
Review of
Aligning Musical Scores with Audio using Hybrid Graphical Models
by
Christopher Raphael

Aligning Musical Scores with Audio using Hybrid Graphical Models takes on the problem of matching notes in musical score with onset times in a live performance. Christopher Rafael's proposed algorithm is based on hidden Markov models. He represents the score as a sequence of chords, each of which has an unknown onset time dependent on tempo and local randomness in the performance. Those onset times and the associated tempo then become the unobserved variables in the Markov model. The observable variables are the power spectra of the recorded performance. The paper presents some informal results as well as a number of potential applications.

The alignment problem taken on in this paper may seem trivial. After all, the score represents a sequence of chords along with the time, in beats, that they should be played. Given only the tempo in beats per minute, one should apparently be able to determine the onset time of each chord. This, however, is not the case. In an expressive performance, tempo changes as a function of time and human error can cause a note to be played slightly before or after the time dictated by the current tempo. Without taking these factors into account, it is impossible to correctly align a real-world performance with a score.

These two variables, tempo variation and local randomization, are the basis for the author's model. Assuming that each variable adheres to a Gaussian distribution with known mean and variance, it is possible to derive the probability of any sequence of onset time. The author further assumes that the power spectrum of any small time slice of a recording will consist primarily of harmonics of the notes that make up the current chord. This assumption refines the probability model by considering the likelihood that a given time-slice corresponds to a given chord in the score. The higher the overall probability of a given sequence of onset times, the more likely that sequence represents a correct estimate of the alignment.

While it is possible to search through all possible alignments to find the one with maximum probability, such an exhaustive search space would be intractably large. The author's solution to this problem is an elegant system of trimming that vastly decreases the number of cases considered while still guaranteeing an optimal result. The idea is that given an subset of chord onsets up to a given time in the performance, all of which end on the same chord, only the sequence with maximum probability up to that point can be a subset of the overall optimal solution. This kind of trimming is closely related to the Viterbi algorithm, one of the most common methods of solving hidden Markov models. The paper presents results showing that trimming reduces the size of the search space by several orders of magnitude.

The paper cites a number of compelling potential uses for this system. One application might music education. The ability to watch the notes of a score highlight while a piece plays would no doubt be useful for students. The paper also suggests the creation a tool that allows a user to fast-forward to a section of a score and listen to playback from that point. The ability to navigate a recording by score rather than just time would be invaluable in a wide variety of fields.

Overall, the proposed algorithm seems very well formulated and is founded on solid theoretical principles. The underlying math is complicated, but is presented clearly enough that the system would be relatively easy to implement. Although few concrete examples on real-world recordings are cited, further testing would be easy and may help refine the system.