

Automatic Extraction of Tempo and Beat from Expressive Performances
Author: Simon Dixon

In this paper Simon Dixon explains his strategy for the beat detection process. His system is a Linux based application that takes as input audio files or MIDI style data and creates predictions for beat time and tempo based on salience of rhythmic events. Because the system uses no top-down musical knowledge to make its decisions it is more suited for multiple genres. The system uses a two pass method that although finishes in less time than the song, cannot be used in real time because of its use of non-causal data.

When using audio data the system applies a high-pass filter and then considers the derivative of the amplitude envelope over a given period. Peaks in this data set are considered onsets. Any onset with sufficient “rhythmic salience” is considered a “rhythmic event”. For the MIDI data case rhythmic events are essentially presented by inspection. In either case, these rhythmic events are used to derive the tempo and beat information.

In the first stage tempo is approximated by clustering the rhythmic events into all possible local combinations. The combinations which can extend across multiple events can allow the system to reduce the weight of events that are uncorrelated to the beat. The clusters are used to interpolate additional beat points. If as a result two clusters have merged in tempo, one is removed. This data is filtered into a ranked list of possible tempo solutions. The best cluster is the one containing the greatest number of events. Ideally a cluster containing all rhythmic events would be the tempo of the smallest metrical interval present in the music. At any given time the best ranking cluster is considered the inter-beat interval or the time between successive beats.

The idea of tempo only indicates the beat rate or frequency. To calculate the phase of the beat or the actual beat time a second phase of beat tracking is used. For each tempo hypothesis created in the first phase, the system instantiates a series of agents to track the song at that rate. The agents are each created under different assumptions of phase for that tempo based on the first five seconds of song time. In most cases one of the agents will start with the right tempo and phase for the actual beat. While running in the main loop correct predictions (within 25ms) of the next event by an agent are rewarded with a rating. Each rating is a factor between 0.5 and 1 based on the salience of the current event. So each agent carries a history, state of current predictions and a rating. As before the agent with the best rate is considered to be the correct beat and reported as the result. When two agents converge into the same beat, one is removed leaving the agent with the better rating.

The results of the system were actually impressive with high ratings for most songs tested. However, it is important to note two things. The first is that results are based on a comparison to a subjective solution by the artist or score and may differ in the actual performance. Second the acceptable tolerance for error to be correct is 25mS. This time is well within the “Just noticeable difference” range for temporal separation. So it is hard to say conclusively the results are accurate.

The biggest draw back to this system is that it cannot be completed in real time. This ultimately cuts down on the available applications of such beat detecting technologies. In addition, the systems reliance on self-inducing “goodness” or “salience” prevents relatively short tempo changes from being factored. In this case although it is robust to syncopations intended beat drops or tempo changes may go unaccounted for. This is because a significant change or drop may be equally wrong for all agents, when this happens the strongest agent would continue to lead in the wrong direction. Ruling out use for some emerging popular music and techno where beat tracking is a desired function for such things as lighting and other events. Another fault would exist when a song has a significant lead-in. Because, the system relies on the first 5 seconds to tune the agents initial predictions a significant lead in might damage the predictions to the point they are unrecoverable. Examples might be Ani DiFranco’s use of orchestral pieces to open a faster paced song, or U2 using long ambient intros in the POP and zooropa albums with relatively low dynamic range resulting in few detected “rhythmic events”.