TEMPO ESTIMATION AND CAUSAL BEAT TRACKING USING ENSEMBLE LEARNING

Michelle L. Daniels
University of California San Diego
Music Department
michelledaniels@ucsd.edu

ABSTRACT

This paper describes submissions to the MIREX 2014 Audio Tempo Estimation and Audio Beat Tracking tasks based on the same causal beat tracking algorithm. For beat tracking, multiple trackers are utilized in an ensemble learning framework to explore different hypotheses about the current tempo and next predicted beat location. At each point in time, the outputs of these trackers are combined using a weighted majority vote, with weights defined as the confidence of a particular tracker in its hypothesis. This provides a single winning tempo and beat phase hypothesis, which is used to determine beat locations for the Audio Beat Tracking task. Intermediate tempo estimates and confidence levels for each tracker are accumulated during this causal beat tracking process and are used to determine a winning tempo estimate in a post-processing stage for the Audio Tempo Estimation task.

1. INTRODUCTION

Audio beat tracking is an inherently ambiguous problem. Ask a group of human listeners to tap their feet in time with a piece of music, and you are likely to find listeners tapping at different tempos (different metrical levels), and perhaps out of phase with each other (some on the off-beat). Machine listeners suffer from the same problem. In addition, different beat tracking algorithms might perform better or worse depending on the content of the music being tracked [9] or the evaluation metric used [2]. Some features work well in music with strong and sharp onsets for example, while others are better for softer attacks. When estimating tempo, different techniques might emphasize faster or slower tempos. This makes it difficult to design a single beat tracking algorithm which can perform well on all genres of music and various tempo ranges.

Ensemble learning methods from the field of machine learning are intended to address this kind of situation. In them, the output of multiple classifiers or regressors is combined to produce results which are better than those of a single classifier. The various classifiers might be trained on different datasets or use different algorithms to determine their results, so that they perform differently depending on the current input. Some algorithms will provide similar or identical results, and those results can be combined in a variety of ways to produce the output of the ensemble.

It has been shown that the task of tempo estimation can benefit from combining the results of multiple tempo estimation algorithms using a voting scheme [7]. Similarly, it should be possible to use an ensemble of beat tracking algorithms to produce a more reliable output than a single algorithm would obtain across a variety of input data. In offline beat tracking systems, which are typically non-causal, the output is a series of estimated beat locations, and it is difficult to see how several such series could be combined to produce a single winning series from an ensemble. However, in online systems, which make decisions about the current tempo and next predicted beat location in a causal fashion, it is possible to examine the output of an ensemble of trackers at successive instants in time, and at each time, combine the results of the trackers to produce the current winning tempo and predicted beat location. Goto's agent-based beat tracking system [6] could in some ways be viewed as such an ensemble learning-based approach, using an ensemble of agents.

The causal beat tracking system used in this submission employs an ensemble of trackers, and is designed to enable comparisons between a variable number of trackers using different approaches, as well as examining different ways to seek consensus among trackers. Each tracker uses different parameters, including input features, approaches to periodicity estimation, and allowed tempo ranges, and the outputs of the trackers are combined after each analysis frame to determine the current winning tempo and estimated beat phase. From this information, the system can predict future beat locations and decide when a beat has occurred for the Audio Beat Tracking Task.

When a constant-tempo input is assumed as in the MIREX Audio Tempo Estimation task, there are many ways in which the winning hypotheses from the causal beat tracking process can be combined to produce the two most salient tempo estimates. For example, a series of inter-beat intervals (IBIs) can be computed from the predicted beat locations generated by the causal system, and these IBIs can be used to estimate the tempo period as several algorithms did in [7]. Alternatively, the winning beat locations themselves could be ignored, and the winning tempo estimates from

the causal system could be grouped to determine the most commonly-chosen tempos. However, these approaches do not take into account the information that can be derived from including all hypotheses from each time instant, not just winning hypotheses. The Audio Tempo Estimation submission therefore uses all intermediate hypotheses from all time instants to determine the best tempo hypotheses for each excerpt.

2. ALGORITHM

2.1 Individual Trackers

Each tracker extracts one feature from the input wavefile. Based on this feature, a periodicity function is computed, from which the current tempo is estimated. This tempo estimate is used to determine the optimal beat phase.

2.1.1 Feature Extraction

Currently, possible features are limited to the following variations on onset detection functions described in [1] and [4]: L1 Magnitude, L1 Magnitude Rectified, L2 Magnitude, L2 Magnitude Rectified, L1 High Frequency Content, L2 High Frequency Content, Complex Domain, L1 Phase Deviation, and L2 Phase Deviation. However, other features, especially ones that might be more indicative of harmonic/chord changes will be added in the future. In this submission, features are computed using a frame size of 1024 samples with a hop size of 512 samples.

2.1.2 Periodicity Estimation

Possible periodicity estimation methods currently include comb filters [8], DFT [11], and auto-correlation [5] [10]. Half of the systems estimate periodicities over 5-second time ranges, and the other half use 10-second time ranges. From its computed periodicity function, each tracker chooses the maximum value in its assigned tempo range as its current tempo estimate. In this submission, four possible overlapping tempo ranges are used: $[30.0, 120.0]$, $[75.0, 165.0]$, $[120.0, 200.0]$, and $[165.0, 280.0]$. Because of the overlap, the highest and lowest possible tempos are covered by fewer trackers, resulting in an inherent bias in the system towards moderate tempos.

2.1.3 Beat Phase Estimation

Each tracker estimates a beat phase in the range $[0.0, 1.0]$, where 0.0 means that a beat is occurring at the current time, and 1.0 represents one tempo period in the future. By correlating a simulated beat sequence with its input feature, the sequence offset producing the maximum correlation is chosen as the most likely alignment.

2.1.4 Tracker Confidence

Each tracker is assigned a confidence value, which a combination of its confidence in its tempo hypothesis and confidence in its beat phase hypothesis.

The tempo confidence value is a combination of four factors. The first factor is a measure of the peakiness of the periodicity function (the relationship between the value of the periodicity function at the chosen tempo period and the mean of the periodicity function). The second factor is continuity with previous tempo estimates made by the same tracker, and the third factor is continuity with the most recent "winning" tempo estimate from the ensemble. The tracker's current confidence is computed by combining a weighted sum of these three factors with a percentage of the tracker's previous confidence.

The beat confidence value is based on the difference between the tracker's next predicted beat location and a multiple of the tracker's chosen tempo period added to the last recorded beat.

2.2 Ensemble of Trackers

2.2.1 Combining Hypotheses

The output of each tracker is a set of three values: the tracker's current tempo estimate, beat phase, and confidence. The tracker's next predicted beat location can be computed from its tempo estimate and beat phase. Pairs of [tempo, next predicted beat location] estimates are then clustered to determine groupings of similar hypotheses. Each cluster is then assigned a score based on the sum of the confidences of each tracker in that cluster. The centroid of the cluster with the largest score is then chosen as the winning hypothesis. This is equivalent to the weighted majority vote used in ensemble learning to combine the output of multiple classifiers [12]: the trackers themselves can be viewed as classifiers, where the classes are defined by quantized [tempo, beat location] pairs and the quantization is performed by the clustering algorithm.

2.2.2 Identifying Beat Locations

Each time the beat phase wraps, a beat location is recorded. The actual beat time is interpolated between the current and previous times, based on when the phase actually reached a value of 1.0.

2.3 Tempo Estimation

Intermediate tempo hypotheses and confidence levels for each tracker at each time instant are accumulated during the causal beat tracking process and are used to determine a single winning tempo estimate in a post-processing stage for the Audio Tempo Estimation task. The tempo hypotheses are clustered using K-means, quantizing hypotheses into discrete tempo classes. The value of K is experimentally chosen to be 10, and the cluster centroids are initialized uniformly across the range of valid tempos.

The confidence levels for each hypothesis in a tempo cluster are accumulated to provide a single score for each cluster. Cluster scores are then adjusted to give greater weight to tempos in integer ratio relationships to other tempos, as in the initialization of Dixon's BeatRoot system [3], but with greater weight given to the smaller tempo in each pair, which helps to compensate for a bias in the system towards tracking at faster metrical levels. Finally, the centroid of the highest-scoring cluster is chosen as the

winning tempo, with the centroid of the second-highest-scoring cluster chosen as the second-place tempo. The relative scores of the two highest scoring clusters are used to determine the relative strength of each hypothesis.

3. IMPLEMENTATION

The causal beat-tracking system used in this submission is implemented in C++ as a class which can be used from any of a number of other applications, including Vamp plugins, Pd externals, or a stand-alone GUI application. For the purposes of each submission, the beat tracker is wrapped in simple cross-platform command-line applications (for Windows, Mac OS X, or Linux) which depend on libsndfile [1] for wavefile reading. The beat tracker itself depends only on the single-precision version of FFTW [2], which it uses for DFT computations.

4. REFERENCES

[1] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler. A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, September 2005.

[2] Matthew E. P. Davies, Norberto Degara, and Mark D. Plumbley. Evaluation Methods for Musical Audio Beat Tracking Algorithms. Technical report, Queen Mary University of London, 2009.

[3] Simon Dixon. Automatic Extraction of Tempo and Beat From Expressive Performances. *Journal of New Music Research*, 30(1):39–58, March 2001.

[4] Simon Dixon. Onset Detection Revisited. In *Proc. of the 9th Int. Conference on Digital Audio Effects, DAFx-06*, pages 133–137, Montreal, Canada, September 2006.

[5] Daniel P. W. Ellis. Beat Tracking by Dynamic Programming. *Journal of New Music Research*, 36(1):51–60, March 2007.

[6] Masataka Goto and Yoichi Muraoka. Beat Tracking based on Multiple-agent Architecture - A Real-time Beat Tracking System for Audio Signals. In *Proceedings of the Second International Conference on Multi-agent Systems*, pages 103–110, 1996.

[7] Fabien Gouyon, Anssi P. Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An Experimental Comparison of Audio Tempo Induction Algorithms. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1832–1844, September 2006.

[8] Anssi P. Klapuri, Antti J. Eronen, and Jaakko T. Astola. Analysis of the Meter of Acoustic Musical Signals. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):342–355, January 2006.

[9] Martin F. McKinney, Dirk Moelants, Matthew E. P. Davies, and Anssi P. Klapuri. Evaluation of Audio Beat Tracking and Music Tempo Extraction Algorithms. *Journal of New Music Research*, 36(1):1–16, March 2007.

[10] Joao Lobato Oliveira, Matthew E. P. Davies, Fabien Gouyon, and Luis Paulo Reis. Beat Tracking for Multiple Applications: A Multi-Agent System Architecture With State Recovery. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):2696–2706, December 2012.

[11] Geoffroy Peeters. Template-Based Estimation of Time-Varying Tempo. *EURASIP Journal on Advances in Signal Processing*, 2007:1–15, 2007.

[12] Robi Polikar. Ensemble Based Systems in Decision Making. *IEEE Circuits and Systems Magazine*, (Third Quarter):21–45, 2006.

[1] http://www.mega-nerd.com/libsndfile/

[2] http://www.fftw.org/