An Empirical Comparison of Tempo Trackers

Simon Dixon
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria
simon@oefai.at

An Empirical Comparison of Tempo Trackers

Simon Dixon
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria
simon@oefai.at

Abstract

One of the difficulties with assessing tempo or beat tracking systems is that there is no standard corpus of data on which they can be tested. This situation is partly because the choice of data set often depends on the goals of the system, which might be, for example, automatic transcription, computer accompaniment of a human performer, or the analysis of expressive timing in musical performance. Without standard test data, there is the risk of overfitting a system to the data on which it is tested, and developing a system which is not suitable for use outside a very limited musical domain. In this paper, we use a large, publicly available set of performances of two Beatles songs recorded on a Yamaha Disklavier in order to compare two models of tempo tracking: a probabilistic model which uses a Kalman filter to estimate tempo and beat times, and a tempo tracker based on a multi-agent search strategy. Both models perform extremely well on the test data, with the multi-agent search achieving marginally better results. We propose two simple measures of tempo tracking difficulty, and argue that a broader set of test data is required for comprehensive testing of tempo tracking systems.

1 Introduction

Much music has as its rhythmic basis a series of pulses, spaced approximately equally in time, relative to which the timing of all musical events can be described. This phenomenon is called the *beat*, and the individual pulses are also called beats. Human subjects are capable of finding the beat with minimal musical training; clapping or foot-tapping in time with a piece of music is not considered a remarkable skill. However, as with many primitive tasks of which humans are capable with apparently little cognitive effort, attempts to model human behaviour algorithmically and reproduce it in computer software have met with limited success.

Various models for extracting the beat from performance data have been proposed. Some are models of human perception, and thus try to mimic human behaviour; others are more goal-oriented, such as those designed for automatic accompaniment, score extraction or automatic transcription. There are also distinctions between on-line and off-line algorithms. Automatic accompaniment systems necessarily use on-line algorithms, as do perceptual models, which might also model the influence of listeners' expectations. Transcription and analysis software tends to process data off-line, as there is the computational advantage that rhythmically ambiguous sections can often be resolved by the context provided by subsequent musical events.

In this paper we compare the statistical approach of Cemgil, Kappen, Desain & Honing (2000; 2001) with the multiple agent approach of Dixon (2000); Dixon & Cambouropoulos

(2000); Dixon (2001). Cemgil, Kappen, Desain & Honing use a Bayesian framework in which the tempo is estimated by a Kalman filter. This allows them to formulate on-line and off-line implementations, and they present in their results that the implementations correctly find approximately 90% of the beats on the Beatles data, with the on-line version performing only marginally worse than the off-line. The beat tracker of Dixon uses a two-stage off-line process: the first finds the tempo of the music (*tempo induction*), and the second synchronises a pulse sequence with the music (*beat tracking*). In each of these stages there may exist multiple hypotheses; these are modelled by a multiple agent architecture in which agents representing each hypothesis compete and cooperate in order to find the best fitting beat sequence. This work also reports results of approximately 90% correctness, on two entirely different data sets, a small set of audio data consisting of jazz and popular songs, and a large set of MIDI format performances of Mozart piano sonatas. The current paper tests this latter system with the data used by Cemgil, Kappen, Desain & Honing (2001), in order to compare the tempo tracking of the two systems, providing the first published cross-validation of tempo tracking systems. Neither of the systems are named; for convenience we will refer to the two systems as the probabilistic tempo tracker (Cemgil, Kappen, Desain & Honing 2001) and the multi-agent tempo tracker (Dixon 2001) respectively.

The content of the paper is as follows: after a brief review of the tempo tracking literature, we describe the architecture and algorithms of the two tempo tracking systems evaluated in this paper. We then present the results of the tests, and conclude with a discussion of the results and the scope of the experiments.

2 Review of Tempo Tracking Research

Early models of rhythmic perception (Steedman 1977; Longuet-Higgins & Lee 1982; Lerdahl & Jackendoff 1983; Povel & Essens 1985), as reviewed and compared by Lee (1991), use musical score data as input, rather than performance data, and take advantage of the simple integer ratios relating the time intervals between note onsets. These methods do not specify how to deal with non-metrical time, that is the expressive and random variations which occur in musical performance data, so we will not consider them further here.

Models of tempo tracking of performance data can be separated into those that use audio data and those that use higher-level symbolic descriptions of the performance (such as MIDI, which explicitly represents note onset times, pitches and relative amplitudes). The audio-based systems tend to have a preprocessing step which extracts symbolic MIDI-like information, or at least the onset times of notes, from the audio data. Subsequent processing is then similar to the MIDI-based systems. The extraction of symbolic information from audio data cannot be performed reliably; that is, there is no known algorithm which can accurately extract the onset times, pitches and amplitudes of notes in an audio signal containing polyphonic music. In fact, it is easily shown that the task is in the general case impossible. However, the tempo tracking performance of humans and recent computer systems in the absence of such accurate symbolic data, demonstrates that such a high level of detail is not necessary for tempo tracking purposes (Dixon 2000). We first describe the systems based on symbolic data, then those using audio input, and finally the probabilistic and multi-agent systems which are compared in this paper.

2.1 Tempo Tracking of Symbolic Performance Data

Systems using symbolic input data usually have access to high level information apart from note onset times (for example pitch and amplitude information), but most tempo tracking algorithms do not use this information. Rosenthal (1992) presents a method for rhythmic parsing of performed melodies, attempting to assign a metrical hierarchy that best matches the input data. The system is usually able to find a correct parse when the performance fulfills the assumptions that there is very little syncopation and no long notes or pauses in the performance.

Tanguiane (1993) presents an information-theoretic approach to rhythmic parsing, which chooses from the possible rhythmic interpretations the one with lowest complexity, in the sense of Kolmogorov. He argues that this is a good functional approximation of human perception. Similarly, Allen & Dannenberg (1990) describe a system that examines multiple hypotheses using beam search with a heuristic that prefers simple musical structures, although their measure of simplicity is not defined.

Desain & Honing (1989) and Longuet-Higgins (1987) present two different methods of quantisation, a connectionist and a symbolic method respectively, which are compared in Desain (1993). The outcome of the paper is the claim that a rhythmic pattern induces an expectation of particular times at which future events should occur, and therefore that a rhythmic parsing system should model these expectations explicitly, in order to provide a contextual framework in which future events can be interpreted with greater certainty.

An alternative approach is to model the expectations implicitly using a nonlinear oscillator (Large & Kolen 1994; Large 1995; 1996). In this work, the initial tempo and beat time are assumed to be known, and the system then tracks the tempo variations using a feedback loop to adjust the oscillator frequency.

2.2 Tempo Tracking of Audio Performance Data

Goto & Muraoka (1995; 1999) describe two systems for beat tracking of audio in real time, the first based on detecting drum sounds and matching to pre-stored patterns, and the second system based on the recognition of harmonic changes, which are assumed to occur at metrically strong locations in time. The systems have a multiple agent architecture with a fixed number of agents implementing various beat tracking strategies via different parameter settings. These systems perform well in the domains for which they were designed, that is, popular music in 4/4 time, with a tempo range of 61–120 beats per minute, and either drum or harmonic rhythms matching the assumed patterns.

A more general approach is taken by Scheirer (1998), who uses a bank of comb filters representing a discrete scale of 150 possible tempos, and passes a heavily processed audio signal through the filterbank, choosing the filter with greatest resonance as the current tempo. This approach is not limited to any particular musical style, however it does not work well for continuous changes of tempo, since the system must repeatedly switch between discrete filters.

2.3 The Probabilistic System

A more principled approach is put forward by Cemgil, Kappen, Desain & Honing (2000; 2001), who model tempo tracking in a probabilistic (Bayesian) framework. The beat times are modelled as a dynamical system with variables representing the rate and phase of the beat, and corresponding to a perfect metronome corrupted by Gaussian noise. A Kalman filter is then used to estimate the unknown variables. To ensure the beat rate is positive, a logarithmic space is used for this variable, which also corresponds better to human perception of time.

In a performance where the system does not have access to the musical score, the beats are not directly observable. Therefore the beats must be induced from the data, which is done by calculating a probability distribution for possible interpretations of performances, based on the infinite impulse response comb filters used by Scheirer (1998).

The parameters for this system are estimated by training on a data set for which the correct beat times are known. From the set of performances of the two Beatles songs, the performances of Michelle were used for training the system, and the performances of Yesterday were used for testing.

2.4 The Multi-Agent System

Dixon (2000) describes an audio beat tracking system using multiple identical agents, each of which represents a hypothesis of the current tempo and synchronisation (phase) of the beat. The system works well for popular music, where tempo variations are minimal, but does not perform well with larger tempo changes. Dixon & Cambouropoulos (2000) extend this work to cater for significant tempo variations as found in expressive performances of classical music. They use the duration, amplitude and pitch information available in MIDI data to estimate the relative rhythmic salience (importance) of notes, and prefer that beats coincide with the onsets of strong notes. In this paper, the salience calculation is modified to ignore note durations because they are not correctly recorded in the data.

Processing is performed in two stages: tempo induction is performed by clustering of the time intervals between near note onsets, to generate the initial tempo hypotheses, which are fed into the second stage, beat tracking, which searches for sequences of events which support the given tempo hypothesis. The search is performed by agents which each represent a hypothesised tempo and beat phase, and try to match their predictions to the incoming data. The nearness of the match is used to evaluate the quality of the agents' beat tracking, and the discrepancies are used to update the agents' hypotheses. Multiple reasonable paths of action result in new agents being created, and agents are destroyed when they duplicate each others' work or are continuously unable to match their predictions to the data. The agent with the highest final score is selected, and its sequence of beat times becomes the solution.

3 Data, Evaluation and Results

The data used in this experiment consists of arrangements of two Beatles songs (*Michelle* and *Yesterday*) performed by 12 pianists (4 professional jazz, 4 professional classical and 4 amateur classical) at each of 3 tempo conditions (slow, normal, fast; as judged by the performer), 3 times for each condition. This gives a total of 2 * 12 * 3 * 3 = 216 performances, plus an additional rendition of Yesterday from memory by 3 of the pianists makes a total of 219 performances.

The evaluation procedure of Cemgil, Kappen, Desain & Honing (2001), which rates the similarity of two sequences of beat times as a percentage, is used to compare the output of the tempo trackers with the beats annotated in the score. If the two sequences are $S_i = s_1, s_2, ..., s_I$ and $T_j = t_1, t_2, ..., t_J$, then the score for the closeness of two beat times s_m and t_n is given by the Gaussian window function $W(d) = exp(-d^2/2\sigma_e^2)$, where $d = s_m - t_n$ and $\sigma_e = 0.040$ sec

Table 1: Average tempo tracking performance ρ, by subject group, tempo condition, and average over all performances.

	Yesterday Data Set		Michelle Data Set
	Probabilistic	Multi-agent	Multi-agent
By subject group			
Professional jazz	95 ± 3	95 ± 7	95 ± 2
Amateur classical	92 ± 8	93 ± 13	94 ± 4
Professional classical	89 ± 7	95 ± 4	88 ± 12
By tempo condition			
Fast	94 ± 5	96 ± 3	94 ± 8
Normal	92 ± 8	95 ± 9	93 ± 6
Slow	90 ± 7	92 ± 12	91 ± 9
Average	91 ± 7	94 ± 9	93 ± 8

Table 2: Average tempo tracking performance ρ for the multi-agent beat tracker performing tempo induction. The low value for the Yesterday data set is due to tracking being performed at the eighth note level in 94 out of 111 cases.

	Multi-agent System:
Data set	Tempo induction & tracking
Michelle	92 ± 5
Yesterday	66 ± 9
Average	79 ± 15

is the width of the window. Then the similarity function is given by:

$$\rho(S,T) = \frac{\Sigma_i \max_j W(s_i - t_j)}{(I+J)/2} * 100 \tag{1}$$

The results comparing the two tempo trackers are shown in Table 1. To provide an equivalent comparison of the two systems, the multi-agent system was initialised with the initial beat time and tempo, replacing the normal beat induction stage performed by this system. The probabilistic system used the Michelle data set for training, so it has test results for only one data set.

To test the tempo induction of the multi-agent beat tracker, the system was also run normally (with no initialisation of tempo or beat time). The quarter note level and the eighth note level were the two metrical levels at which beat tracking was most successful. On the Michelle data, the quarter note level was chosen in 107 of 108 cases, the eighth note level being chosen in the remaining case. With the Yesterday data, the quarter note level was chosen in 15 of the 111 cases, with 94 cases choosing the eighth note level. The remaining 2 cases were close to the quarter note level, but outside the tolerance of 10%. Without adjusting for the use of different metrical levels, the beat tracking evaluation results are lower, as only every second eighth note beat matches a quarter note beat (see Table 2).

4 Discussion

On the given test data, we have shown that the system of Dixon has better average case tempo tracking performance than the system of Cemgil, Kappen, Desain & Honing, although the differences are not generally significant. However, despite the size of the test set, it is not clear that it adequately tests tempo tracking in general, as both pieces are quite simple arrangements with a very clear beat.

The difficulty of tempo tracking of performance data has two components – the difficulty due to the rhythmic complexity of the piece, which can be estimated from the musical score, and the difficulty due to expressive and random variations in tempo and timing introduced by the performer. As a rough estimate of the former, the proportion of beats on which no event occurs and the proportion of events which do not occur on beats are the simplest indicators. We combine these in equation 2 to compute a *rhythmic complexity index* (RCI):

$$\mathrm{RCI} = \frac{UnmatchedBeats + UnmatchedEvents}{Matches + UnmatchedBeats + UnmatchedEvents} \tag{2}$$

where $Matches$ is the number of beats on which an event occurs, $UnmatchedBeats$ is the number of beats on which no event occurs, and $UnmatchedEvents$ is the number of events which do not fall on a beat. The RCI lies between 0 (for an isochronous sequence) and 1 (for a sequence where no event corresponds to a beat). This gives a better measure of tempo tracking complexity than the C-score of Povel & Essens (1985), which is a weighted sum of the number of beats with unaccented events and the number of beats with no event, but is not normalised, so that repetitions of segments increase the estimated complexity of a piece. Other complexity measures, such as those discussed by Shmulevich, Yli-Harja, Coyle, Povel & Lemström (2001), are more relevant to data compression than to beat tracking difficulty.

Applying the rhythmic complexity index to the Michelle score, we note that the score contains 105 beats, with only 4 of these having no event occurring upon them, and a further 22 events which occur between beats, giving an RCI of 0.20. (These figures vary slightly between performances.) This means that a naive algorithm which decides that every event is a beat and that there are no other beats, will score at least 101 / ((105 + 123) / 2) = 88.6%!

Variations in timing and tempo, depending on their extremity, can cause problems for tempo tracking systems, particularly for those using on-line algorithms. With large changes, it becomes impossible for a system lacking the score to distinguish between a change in a rhythmic pattern and a change in performance timing. The standard deviation of the inter beat intervals can be used as a simple indicator of the difficulty of tempo tracking due to performance timing variations. For the Beatles data set, the average standard deviation in beat intervals was 0.058 seconds, or approximately 10% of the average beat interval. Variation increased with beat interval (remaining at approximately 10% of the beat interval for the various tempo conditions), which explains the lower tempo tracking scores for slower performances, since the evaluation function has a constant window width of 0.04 seconds.

As a comparison, we take the first movement of Mozart's Piano Sonata in C major (KV279), played by a professional classical pianist. The score contains 400 beats, 2 of which have no event played on them, but there are 1052 events which occur off the beat at defined score times, plus another 113 notes which have no defined score time (for example trills, ornaments), giving an RCI of 0.75. That is, the rhythmic complexity of the Mozart score is much higher than for the Beatles arrangements, as expected. On the other hand, the standard deviation of the beat intervals is also very close to 10% of the beat interval, so the difficulty due to timing variations seems to be equivalent. In order to further advance tempo tracking research, this sort of comparitive study should be extended to include such complex real world data sets, as used by Dixon & Cambouropoulos (2000).

Another interesting study would be to compare the nature of the errors made by the two systems, to attempt to isolate the "difficult" cases for tempo tracking. A comparison with human tapping would also be enlightening in this respect.

Acknowledgements

This research is part of the project Y99-INF, sponsored by the Austrian Federal Ministry of Education, Science and Culture (BMBWK) in the form of a START Research Prize. The BMBWK also provides financial support to the Austrian Research Institute for Artificial Intelligence. The Beatles data was collected by the Music, Mind, Machine Group, NICI, University of Nijmegen (for more details, see (Cemgil, Kappen, Desain & Honing 2001) or the MMM web site `http://www.nici.kun.nl/mmm`). Thanks also to Emilios Cambouropoulos, Werner Goebl and Gerhard Widmer for many helpful discussions about tempo tracking.

References

Allen, P. & Dannenberg, R. (1990). Tracking musical beats in real time. In *Proceedings of the International Computer Music Conference*, pages 140–143. International Computer Music Association, San Francisco CA.

Cemgil, A., Kappen, B., Desain, P. & Honing, H. (2000). On tempo tracking: Tempogram representation and Kalman filtering. In *Proceedings of the 2000 International Computer Music Conference*, pages 352–355. International Computer Music Association.

Cemgil, A., Kappen, B., Desain, P. & Honing, H. (2001). On tempo tracking: Tempogram representation and Kalman filtering. *Journal of New Music Research*. To appear.

Desain, P. (1993). A connectionist and a traditional AI quantizer: Symbolic versus sub-symbolic models of rhythm perception. *Contemporary Music Review*, 9:239–254.

Desain, P. & Honing, H. (1989). Quantization of musical time: A connectionist approach. *Computer Music Journal*, 13(3):56–66.

Dixon, S. (2000). A lightweight multi-agent musical beat tracking system. In *PRICAI 2000: Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, pages 778–788. Springer.

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*. To appear.

Dixon, S. & Cambouropoulos, E. (2000). Beat tracking with musical knowledge. In *ECAI 2000: Proceedings of the 14th European Conference on Artificial Intelligence*, pages 626–630. IOS Press.

Goto, M. & Muraoka, Y. (1995). A real-time beat tracking system for audio signals. In *Proceedings of the International Computer Music Conference*, pages 171–174. Computer Music Association, San Francisco CA.

Goto, M. & Muraoka, Y. (1999). Real-time beat tracking for drumless audio signals. *Speech Communication*, 27(3–4):331–335.

Large, E. (1995). Beat tracking with a nonlinear oscillator. In *Proceedings of the IJCAI'95 Workshop on Artificial Intelligence and Music*, pages 24–31. International Joint Conference on Artificial Intelligence.

Large, E. (1996). Modelling beat perception with a nonlinear oscillator. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*.

Large, E. & Kolen, J. (1994). Resonance and the perception of musical meter. *Connection Science*, 6:177–208.

Lee, C. (1991). The perception of metrical structure: Experimental evidence and a model. In Howell, P., West, R. & Cross, I., editors, *Representing Musical Structure*, pages 59–127. Academic Press.

Lerdahl, F. & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. MIT Press.

Longuet-Higgins, H. (1987). *Mental Processes*. MIT Press.

Longuet-Higgins, H. & Lee, C. (1982). The perception of musical rhythms. *Perception*, 11:115–128.

Povel, D. & Essens, P. (1985). Perception of temporal patterns. *Music Perception*, 2(4):411–440.

Rosenthal, D. (1992). Emulation of human rhythm perception. *Computer Music Journal*, 16(1):64–76.

Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America*, 103(1):588–601.

Shmulevich, I., Yli-Harja, O., Coyle, E., Povel, D. & Lemström, K. (2001). Perceptual issues in music pattern recognition: Complexity of rhythm and key finding. *Computers and the Humanities*, 35:23–35.

Steedman, M. (1977). The perception of musical rhythm and metre. *Perception*, 6:555–569.

Tanguiane, A. (1993). *Artificial Perception and Music Recognition*. Springer.