BEAT TRACKING WITH A TWO STATE MODEL

M. E. P. Davies and M. D. Plumbley

Queen Mary, University of London
Centre for Digital Music
Mile End Road, London E1 4NS, UK

ABSTRACT

In this paper we apply a two state switching model to the problem of audio based beat tracking. Our analysis is based around the generation and application of adaptively weighted comb filterbank structures to extract beat timing information from the mid-level representation of an input audio signal known as the onset detection function [1]. We evaluate our system using a previously published dataset [2] and in performing a direct comparison with the current state of the art, present comparable results.

1. INTRODUCTION

The principal aim of a beat tracking system is to replicate the human ability of tapping in time to music. Within the field of music information retrieval two important applications for beat tracking can be cited: i) automatic music transcription and ii) automatic musical accompaniment. In the simplest context, a successful algorithm should be able to extract two parameters from an audio signal: a measure of the rate at which beats occur (which we will call the *beat period*) and identifying when they occur (i.e. finding their phase, or *beat alignment*). This seemingly simple problem, at which humans, even without any musical knowledge, seem particularly adept, does not translate so trivially into a computational environment, especially in situations where the tempo of the input varies.

Under certain conditions however, the problem of beat tracking becomes less complicated. This appears directly related to the number of assumptions made about the characteristics of the input signal. Goto's approach [3] is an excellent example. By constraining all input signals to be in 4/4 time and assuming approximately constant tempo falling within the range of 61 and 120 bpm (beats per minute) his system is able to perform very robust beat tracking in real-time. However when attempting the design of a more general beat tracking system, performance naturally decreases. The problem which poses the greatest difficulty is that of tempo variation occurring as a result of natural variation as well intentional timing changes, known as *expressive timing*.

A number of broad approaches to beat tracking have been proposed, including those which employ probabilistic formulations to the problem [4, 2] as well as those more grounded in signal processing techniques such as comb filter analysis [5, 6]. The approach we present borrows from both of these fields. We use combine comb filtering techniques in conjunction with a state space switching model to perform beat tracking using a General and Context Dependent Model.

Matthew Davies is supported by a College Studentship from Queen Mary, University of London.

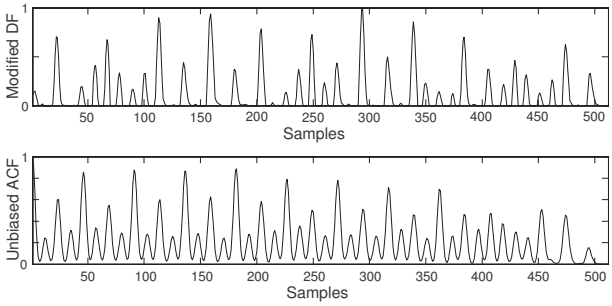

Fig. 1. Modified Detection Function (DF) (top) and Unbiased Autocorrelation Function (ACF) (bottom)

The remainder of this paper is structured as follows. Section 2 presents the General Model for beat tracking followed a description of the Context Dependent Model in Section 3. The combination of the two models is presented in Section 4. Results are given in Section 5 with Conclusions in Section 6.

2. GENERAL MODEL

In this section we describe the general approach taken towards beat tracking as presented in [7]. Beat locations are derived from a two stage process, which begins by identifying the beat period, followed a separate beat alignment stage using the current period estimate. The analysis is frame based, using a window of 512 onset detection function samples (whose resolution is 11.6ms) with a step increment of 128 samples, giving an overlap of 75%.

2.1. Onset Emphasis

The principal input to the system is the mid-level representation of an audio signal known as the *onset detection function* [1]. Primarily used in the extraction of note onset times from audio, the detection function can be considered a continuous representation of onset *emphasis* wherein the local maxima, or peaks of the signal represent the onset locations - a property we shall exploit to identify beat locations.

A number of approaches to generating the detection function exist, where different properties of the audio signal are emphasised. We select the complex domain approach [1] as it is sensitive both to *tonal* and *percussive* events within musical audio signals. The detection function, $df[n]$ of an audio input $x[n]$ is created by measuring the complex spectral difference across all frequency

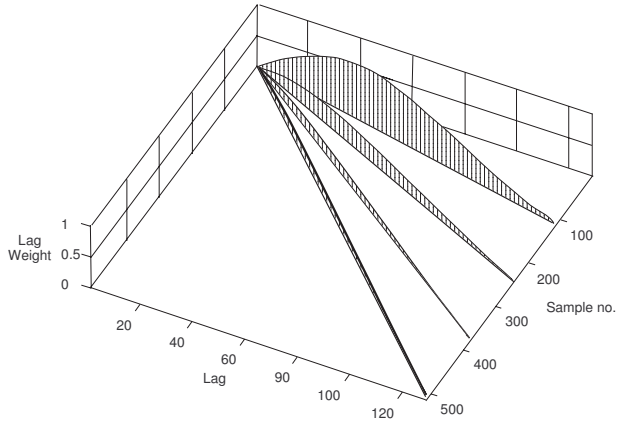

Fig. 2. 3-D view of Lag Weighted Comb Filterbank, M

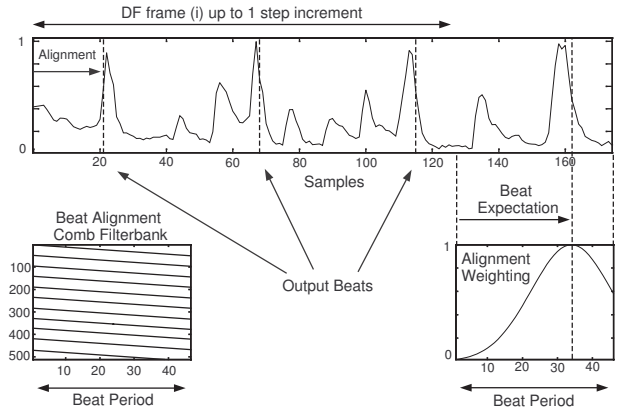

Fig. 3. Beat Placement and Alignment Weighting

bins, k, between a target, $\tilde{X}_k[n]$, and observed, $X_k[n]$, Short Time Fourier Transform (STFT) frame:

$$df[n] = \frac{1}{N}\sum_{k=0}^{N} ||\tilde{X}_k[n] - X_k[n]||^2 \quad (1)$$

The detection function (DF) is then low pass filtered and subjected to an adaptive median threshold, as in [1], to yield a Modified DF (top plot, Fig 1) as the input to the system.

2.2. Beat Period Estimation

The beat period, τ, is found by identifying the most salient lag l from an *unbiased* autocorrelation function (ACF), $\hat{r}_{df}[l]$, of the current detection function frame (of length N)

$$\hat{r}_{df}[l] = ((\sum_{n=0}^{N-1} df[n]df[n-l])(|(l-N)|)) \quad (2)$$

where lag (in DF samples) can be converted to tempo (in bpm) using the relation: $tempo = 60/(l * 0.0116)$, and 11.6ms is the resolution of the DF.

The ACF is passed through a *shift-invariant* comb filterbank, M, covering a lag range of 1 to 128 samples (Fig 2). The elements of M grow wider at each harmonic but are normalized such that each harmonic has equal influence over the location of the beat period. This prevents any metrical bias. Each row of M is scaled by $R_w[l]$, a lag weighting function derived from the Rayleigh Distribution Function, with parameter $b = 43$ (corresponding to the *preferred* tempo of 120 bpm [8])

$$R_w[l] = \frac{l}{b^2} e^{\frac{-l^2}{2b^2}} \quad (3)$$

This weighting, which is similar to the resonance model for beat period [8] acts as a *band pass filter* to encourage the beat period to be extracted in the approximate tempo octave range of 80 to 160 bpm. The inclusion of R_w effectively transforms the goal of the General Model into that of *preferred beat tracking* as signals with tempi outside the effective range of the weighting function will observed at a metrical level within this range.

An accurate value for τ (which overcomes the inverse tempo to lag relationship) is found by finding the column of M which best matches the ACF, i.e. $\arg\max_l(\hat{r}_{df} \times M)$. This column is then used as a *mask* on the ACF with

$$\tau = \frac{1}{P}\sum_{p=1}^{P} \frac{h_p}{p} \quad (4)$$

where P is the number of harmonics in M and h_p is the index of the local maxima of the ACF for the p^{th} harmonic.

2.3. Beat Alignment and Placement

Having found the beat period for the current analysis frame, this value can be used to find the beat alignment, ϕ, defined as the offset from the start of the current frame, t_i, to where the *first* beat should occur.

The method for identifying ϕ is similar to the way in which τ was found, as a comb filter matrix is again used. However instead of the passing the ACF (which by definition is *zero-phase*) through the comb filter, the DF frame is used instead. The properties of the *alignment* matrix comb filter, A (shown in the bottom left of Fig 3) differ from those of M as the elements, which occur at integer multiples of τ allow a search through all possible shifts of τ in the range t_i to $t_i + \tau$ with ϕ found using:

$$\phi = \arg\max_n(df[n] \times A) \quad (5)$$

Beats within the current DF frame can now be placed up to *one* step increment beyond t_i using the following relationship

$$\gamma_m = (t_i + \phi) + (m-1)\tau \quad (6)$$

where γ_m is m^{th} beat in the current frame.

3. CONTEXT DEPENDENT MODEL

When tested in [7] the General Model demonstrated promising results which confirmed the overall approach as appropriate for the problem of beat tracking. However the system's primary failure was in producing a consistent output. The rate at which beats occurred was found to frequently (and unpredictably) switch between metrical levels as well as switching between on-beats to off-beats locations. Given the importance of consistency both in terms of

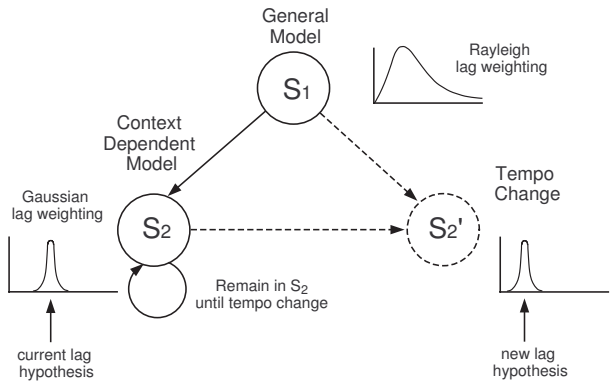

Fig. 4. State Switching Model

possible applications for a beat tracking system, and the emphasis given in evaluation procedures for the longest continuous segment [9, 6, 2] we chose to refine the General Model to make better use of the available rhythmic information. We therefore propose a Context Dependent Model.

3.1. Beat and Measure Period

Within the Context Model we replace the Rayleigh weighting applied to M by a *beat period dependent* Gaussian lag weighting, $G_w[l]$

$$G_w[l] = e^{\frac{-(l-\tau)^2}{2\sigma^2}} \tag{7}$$

where the mean corresponds to an average beat period for the input (which could be derived from a number of individual measurements using the General Model or equally from a notated value, if available). The variance, $\sigma^2 = 3.9$ was empirically derived by analysis of the distribution of *tempo normalized* inter beat intervals from the hand labelled beat database used in this paper. It has an important property in that it allows for some expressive timing variation, but is narrow enough to give a consistent output, even when the input ACF displays very little noticeable periodicity.

In addition to incorporating G_w a simple test is performed to infer the measure level periodicity within the ACF. Currently this involves the classification between *duple* and *triple*, i.e. music that is in either 4/4 or 3/4 time. The decision is made by evaluating the following condition related to energy in the ACF: $(2\tau + 4\tau > 3\tau + 6\tau)$. If true, we assume duple time, else triple time. This then enables the structure of M to be altered such that it consistent with the meter classification. A metrical bias can now be imposed into M by setting the appropriate number of harmonics (either 3 or 4) and scaling them to give most emphasis to detecting measure level periodicity from which the beat periodicity can then be inferred.

3.2. Beat Expectation

The problems observed in phase-switching were a direct result of the *memoryless* state of the General Model. By making independent judgements about beat alignment at every frame, the alignment value typically locked to the strongest peak in the first period of the DF, which could have resulted from an on-beat, off-beat or an accented event unrelated to the beat structure. We aim to resolve this problem, by the generation of an adaptive Gaussian alignment weighting to apply to each row of A, the beat alignment matrix, thus invoking the concept of *beat expectation*. The procedure for generating the alignment weighting is shown in Fig 3. In the General Model, beats were placed only up to one step increment beyond the start of the current analysis frame. However, assuming that the observed beat period will not vary too greatly (a valid assumption given the narrow Gaussian weighting, $G_w[l]$), a beat prediction can be made by identifying the location of the *next* possible beat, i.e. $\gamma_{last} + \tau$ which acts as the mean of the Gaussian alignment weighting. The variance of this weighting was derived to prevent phase-switching, and $\sigma^2 = \tau/4$ was found to be ideal.

Algorithm Used	CML Cont.	CML Total	AML Cont.	AML Total
Davies - GM	26.4	50.4	29.6	59.8
Davies - TSM	53.6	60.8	66.1	81.8
Hainsworth [2]	45.1	52.3	65.5	80.4
Scheirer [5]	23.8	38.9	29.8	48.5
Klapuri [6]	55.9	61.4	71.2	80.9

Table 1. Results comparing different beat tracking algorithms under four conditions: CML - Correct Metrical Level; AML - Allowed Metrical Level; Cont. - Longest Continuous segment, Total - Total number of correct beats

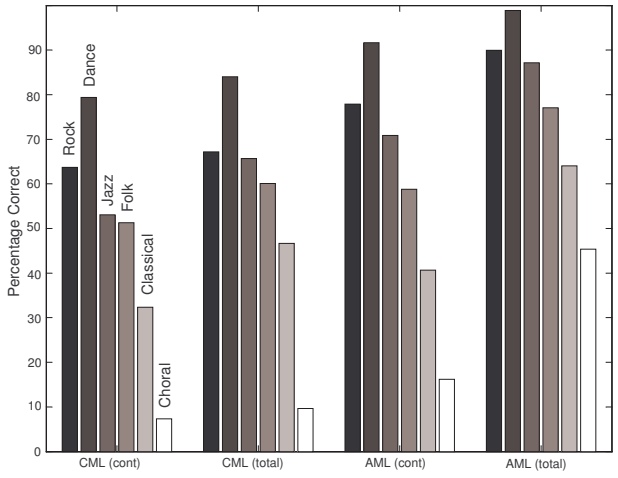

Fig. 5. Comparison of results across musical genre

4. STATE SWITCHING

Having presented both the General and Context Dependent Models we now seek to combine the two, and describe the conditions under which state transitions occur. In the initial instance, there has been no observation of beat period from which to form a Context Model. Therefore beat analysis begins in state S_1 (see Fig 4) using the flat alignment weighting. The Context Model is generated (and hence the transition from S_1 to S_2 occurs) as a result of observing three consecutive *consistent* beat period values such that:

$$\mathrm{abs}(2\tau_\mathrm{r}(\mathrm{i}) - \tau_\mathrm{r}(\mathrm{i}-1) - \tau_\mathrm{r}(\mathrm{i}-2)) < \eta \tag{8}$$

where $\tau_r(i)$ is the beat period using the Rayleigh weighting for frame i, and $\eta = \sigma^2$ from eq. (7).

Beat placement now uses period measurements taken from the Context Model in conjunction with the beat expectation alignment weighting (sec. 3.2). We remain in state S_2 until the observation of a new *consistent* lag hypothesis, defined by:

$$\mathrm{abs}(\tau_\mathrm{r}(\mathrm{i}) - \tau_\mathrm{g}(\mathrm{i})) > \eta \tag{9}$$

where the consistency condition is as in eq. (8) and $\tau_g(i)$ is beat period from the Context Model. Once a new beat period hypothesis has been selected, (i.e. the transition from S_2 to S_2') state S_2 is no longer valid and is removed from the system.

5. RESULTS

We evaluated our algorithm using the same test dataset and performance criteria as in [2]. The database contains 222 audio tracks, each around a minute in length, covering the following range of musical genres: rock/pop (68), dance (40), jazz (40), folk (22), classical (30) and choral (32). The *ground truth* against which the beats were compared, was generated by recording beats clapped by a musician. Each annotated track was then subjectively analysed and any mis-placed beats were manually corrected.

The primary performance metric is defined as the ratio of the *longest continuous correctly tracked* segment to the length of the input signal, and has been used in several beat tracking studies [9, 6, 2]. Beats are considered to have met this criterion if they occur within $\pm 15\%$ of the annotated beat onset time and the beat period is within $\pm 10\%$ of the annotated value. In addition to this rather strict method, a second criterion is defined as the *total number of correct beats*, where the continuity requirement is relaxed. In order to allow for the metrical ambiguity that occurs when humans tap along to music - that in some cases beats are tapped at twice or half the correct level, or on the off-beat instead of the on-beat, results are re-calculated to allow for this behaviour. Therefore to summarize, we use a total of 4 criteria: i) the longest continuous segment, or *Cont.* at the correct metrical level, *CML*; ii) the *total* number of correct beats, at *CML*; iii) the longest continuous segment at the allowed metrical levels *AML* and iv) the *total* number of correct beats at *AML*. Table 1 shows the performance of our *Two State Model* in comparison with the General Model, as well as the implementations of: Hainsworth [2], Scheirer [5] and Klapuri [6]. A breakdown of the performance of our system is shown in Fig 5 for each genre within the test set.

As can be seen from the results in Table 1, a significant improvement has been achieved by the inclusion of the Context Model, such that the Two State Model clearly outperforms the General Model. In terms of a direct comparison with the other published algorithms, the General Model achieves equivalent results to the Scheirer implementation [5], with the Two State Model showing better performance than that of Hainsworth's approach [2], but marginally poorer than that of Klapuri [6], in all categories except the total number of correct beats at *AML*.

The results shown in Fig 5 demonstrate an expected pattern across musical genres with the best performance in the rock/pop and dance categories. The most prominent failure of the algorithm is in addressing choral music, where there is often very little, if any coherent beat structure. The algorithm also relies on the tempo of the input to remain constant within each analysis frame - an assumption which does not hold in cases where there is expressive timing. Our approach to meter analysis is not particularly robust, as it can only address music which is in 3/4 or 4/4 time. The extraction of *downbeat* indices (the first beat of each bar) should also improve performance by adding some higher level structure to the system. The final point to address is that of causality, as the current approach makes use of future data and could not therefore be implemented in real-time, however work towards a causal implementation is underway.

6. CONCLUSIONS

In this paper we have applied a two state switching model to the problem of beat tracking for musical audio signals. The inclusion of a context dependent model has been shown to significantly improve upon the results achieved in our previous work [7], and is now comparable with the current state of the art [2, 6]. Further work is underway to investigate a causal implementation in addition to performing more sophisticated meter analysis.

7. ACKNOWLEDGEMENTS

The authors would like to thank Nick Collins and Steve Hainsworth for providing the test database and results used within this paper.

This research has been partially funded by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents). More information can be found at the project website http://www.semanticaudio.org.

8. REFERENCES

[1] J. P. Bello, C. Duxbury, M. E. Davies, and M. B. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 553–556, July 2004.

[2] S. Hainsworth, *Techniques for the Automated Analysis of Musical Audio*, Ph.D. thesis, Cambridge University, April 2004.

[3] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, pp. 159–171, June 2001.

[4] A. T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempogram representation and Kalman filtering," *Journal Of New Music Research*, vol. 29, no. 4, 2000.

[5] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *Journal of Acoustical Society of America*, vol. 103, pp. 588–601, January 1998.

[6] A. P. Klapuri, "Musical meter estimation and transcription," in *Cambridge Music Signal Processing Colloquium*, 2003.

[7] M. E. P. Davies and M. D. Plumbley, "Causal tempo tracking of audio," in *5th International Symposium on Music Information Retrieval*, October 2004.

[8] L. van Noorden and D. Moelants, "Resonance in the perception of musical pulse," *Journal of New Music Research*, vol. 28, no. 1, pp. 43–66, March 1999.

[9] M. Goto and Y. Muraoka, "Issues in evaluating beat tracking systems," in *Working Notes of the IJCAI-97 Workshop on Issues in AI and Music - Evaluation and Assessment*, 1997.