CAUSAL TEMPO TRACKING OF AUDIO

Matthew E. P. Davies and Mark D. Plumbley
Centre for Digital Music
Queen Mary University of London

ABSTRACT

We introduce a causal approach to tempo tracking for musical audio signals. Our system is designed towards an eventual real-time implementation; requiring minimal high-level knowledge of the musical audio. The tempo tracking system is divided into two sections: an onset analysis stage, used to derive a rhythmically meaningful representation from the input audio, followed by a beat matching algorithm using auto- and cross-correlative methods to generate short term predictions of future beats in the audio. The algorithm is evaluated over a range of musical styles by comparing the predicted output to beats tapped by a musician. An investigation is also presented into three rhythmically complex beat tracking problems, where the tempo is not constant. Preliminary results demonstrate good accuracy for this type of system.

Keywords – Tempo tracking, beat analysis, onset detection, rhythmic analysis

1. INTRODUCTION

In this paper, we describe an approach to causal tempo tracking that could form part of an eventual system for real-time *automatic accompaniment*. The aim of our tempo tracking system is to produce a beat prediction from musical audio in real-time: in effect, to "tap along" to the music.

While beat tracking has perhaps not received the same level of interest as polyphonic pitch tracking for automatic music transcription, there are nevertheless a number of existing approaches to beat tracking and tempo analysis. Some are based on symbolic input such as MIDI: Brown [5], for example, measures the autocorrelation derived from MIDI note onset times to infer metric structure. Raphael [13] uses prior knowledge of the musical piece to perform accurate score following for an accompaniment system based on the observation of a time-varying tempo process. Allen and Dannenberg [1] track beats in real-time using note onset times with a beam search to maintain multiple hypotheses of the tempo at any one time.

The specification of an audio input increases the complexity of tempo tracking due to the pre-processing required to derive a meaningful, *symbolic-like* representation [6] on which to perform the analysis. Depending on the desired application this may be performed offline to compensate for errors introduced in pre-processing. Our motivation is towards a real-time application. A comprehensive review of tempo tracking research, including offline approaches to the problem is given in [6].

The real-time beat tracking system introduced by Scheirer [14] employs an onset analysis stage over six octave frequency bands which is passed to parallel comb filterbanks to extract the beats. This requires a 2-dimensional search over tempo and phase to find the best matching beat position. Goto [10] also describes a real-time beat tracking system using an audio input. Here a range of different drum identification models are used to match an audio input to drum beats on different instruments, these are combined with harmonic information related to chord changes, to infer and track the beat structure.

An ideal beat-tracking system for our eventual aim of an automatic accompaniment system would have a number of particular properties: (i) audio input, rather that symbolic such as MIDI; (ii) causal, such that a beat prediction is produced from knowledge of the "past" signal only; (iii) versatile to changes in tempo and style; (iv) minimal knowledge required about the musical piece to be tracked; and (v) computationally efficient to be implemented in real-time. We therefore set out to design a beat tracking approach that would be suitable for this purpose.

Our approach combines recent work on onset detection [2] with a beat prediction method based on the autocorrelation function (ACF) of the onset detection function. While it has been suggested that the ACF would not be suitable for beat tracking, due to the loss of phase information [14], we combine this with a separate phase matching stage to recover the beat timings. In this way we separate the two-dimensional search for beat-period and beat-alignment (as in [14]) into two one-dimensional searches: one for beat-period, followed by the second for beat-alignment once the beat period is known, hence simplifying the search process.

The paper is structured as follows. In Section 2 we describe the onset detection stage, followed by the tempo estimation in Section 3. Section 4 details the beat alignment and prediction stages followed by results and analysis in Sections 5 and 6. An overall discussion of the system is given in Section 7 with conclusions in Section 8.

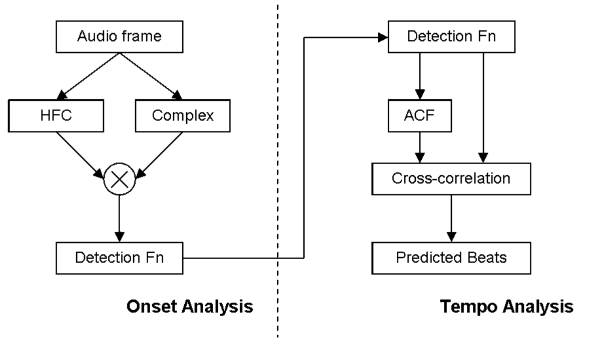

Figure 1. Algorithm Operation Overview. (HFC = high frequency content, ACF = autocorrelation function)

2. ONSET ANALYSIS

The aim of the onset analysis stage is not to explicitly detect the locations of note onsets, rather to generate a mid-level representation of the input which emphasizes the onset locations - similar in principle to the multi-band, envelope extraction processing in [14]. To reflect this need we choose the *onset detection function* [2] - a continuous signal with peaks at onset positions, as the input to the tempo analysis stage. An example detection function is shown in Figure 2.

In our system we use two detection functions - an HFC [12] and a complex domain approach [3] which are multiplied together to create a single input for the tempo analysis stage (as shown in the left hand plot of Figure 1). The combination of these two detection functions has been shown to give improved performance for onset detection than when used individually [4]. Each onset detection function is generated from frame based analysis using a window size of 1024 samples and a step increment of 512 samples, from audio sampled at 44.1kHz, giving a temporal resolution of 11.6 ms. To aid in the development process the two detection functions were generated prior to run time, however both are computationally feasible in real-time [4].

2.1. High Frequency Content (HFC)

Masri and Bateman [12] present an approach to energy based onset detection, using a linear weighting corresponding to bin frequency k of the Short Time Fourier Transform (STFT) frame $X_k[n]$ of audio input $x[n]$ to emphasize the high frequency energy within the signal, giving the detection function output $df_h[n]$ given by:

$$df_h[n] = \sum_{k=0}^{N} k|X_k[n]| \tag{1}$$

This technique is particularly appropriate for emphasizing percussive type onsets, most notably cymbals, where the transient region of the instrument hit is mainly composed of energy at high frequencies.

2.2. Complex Domain Onset Detection

While the HFC approach is suited for signals with strong percussive content, it performs poorly when applied music with non-percussive onsets, such as those created by a bowed violin. We therefore incorporate a second detection function, as implemented by Bello et al [3] that is able to contend with a wider variety of signal types.

$$df_c[n] = \frac{1}{N}\sum_{k=0}^{N} ||\tilde{X}_k[n] - X_k[n]||^2 \tag{2}$$

The complex detection function $df_c[n]$ shown in equation (2) is a combined energy and phase based approach to onset detection. It portrays the complex spectral difference between the current frame $X_k[n]$ of a STFT and a predicted target frame, $\tilde{X}_k[n]$. Detection function peaks are the result either of energy change or deviations in phase acceleration. These deviations occur in transients as well as during pitch changes (often called *tonal onsets*, where no perceptual energy change is observed) enabling the approach to detect onsets in a wider range of signals. A more complete derivation and discussion may be found in [2].

3. TEMPO ANALYSIS

Our tempo analysis process shown in the right of Figure 1, where the beat period is estimated from the autocorrelation function (ACF) of the detection function. A frame based approach is implemented, using a window size of 512 detection function samples, with a step increment of 128 samples (75% overlap), giving an audio analysis frame of approximately 6 seconds.

3.1. Auto-correlation function (ACF)

The first stage in estimating the beat period is to calculate the ACF of the input detection function frame. Given the assumption that the detection function has strong peaks at note onsets and the rhythmic content of the signal is coherent (i.e. that there is some rhythmic structure) the peaks of the ACF will correspond to periodic events within the signal at certain time lags, shown in Figure 2. The ACF, r_{df}, at lag, l, is derived as follows:

$$r_{df}[l] = \sum_{n=0}^{N-1} df[n]df[n-l] \tag{3}$$

where $df[n]$ is the combined detection function, and N is the length of the detection function frame. The lower plot in Figure 2 shows a linear trend down from the start of the ACF. It is necessary to correct this so the analysis is not biased. A modified ACF, $\hat{r}_{df}$, with the bias removed is calculated in equation (4). The signal is squared to emphasize the peaks.

$$\hat{r}_{df}[l] = ((\sum_{n=0}^{N-1} df[n]df[n-l])(1 + |(l-N)|))^2 \tag{4}$$

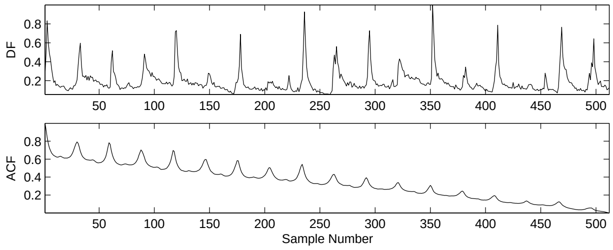

Figure 2. Detection function (upper plot) and corresponding ACF (lower plot)

By selecting the lag of a particular peak in the ACF, an estimate of the tempo can be made by applying equation (5), where β converts the observed ACF lag l into the recognisable units of tempo: beats per minute (bpm).

$$\text{tempo} = \frac{\beta}{l} \tag{5}$$

3.2. Beat Period Estimation

The inverse *tempo-lag* relationship results in poor tempo resolution at short lags, meaning a single peak is often insufficient to obtain an accurate value for the beat period. Figure 2 shows the ACF has rhythmically related peaks at longer lags. In a similar context to Brown's *narrowed autocorrelation function* [5], greater accuracy can be achieved by extracting more than one peak. We combine four related lag observations from the ACF, and take the mean to arrive at a value which compensates for the poor resolution.

An efficient means to extract a number of arithmetically related events from a signal (in this case rhythmically related lags from an ACF) is to use a matrix-based comb filter approach [7]. Because the ACF is zero-phase, the search for the lag corresponding to the beat period is shift-invariant. This permits the ACF to be passed through a single bank of comb filters. Each filter is matched to a different fundamental beat period, covering all lags in the range of 1 to 128 samples, where the maximal lag output τ corresponds to the best matching beat period of the signal frame:

$$\tau = \arg\max_{l}(\hat{r}_{df} \times M) \tag{6}$$

where $\hat{r}_{df}$ is a row vector of length N and M is an $(N \times L)$ matrix, with L as the number of possible lags.

3.3. Lag Weighting Function

The implementation of the comb filterbank, as derived in section 3.2, did not lead to the robust extraction of lags, even in simple test cases. An equal weighting for all lags gave too much freedom to the possible beat period output, allowing lags of just a single sample to be considered in the beat period estimation and resulted in frequent inconsistency of the output. An intuitive solution would be to impose explicit limits on the range of lag values - to have minimum and maximum values defined by a desired range of tempi, derived from equation (5). However, the inverse relationship between lag and tempo means that an equal weighting of lags does not correspond to equal tempo weighting which results in bias towards slower tempi. An alternative is to derive a weighting function for the allowed lag values, imposing *implicit* limits on an acceptable tempo range and reducing the bias introduced from the non-linear *tempo-lag* relationship.

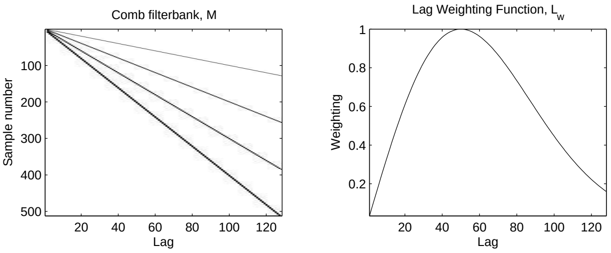

Figure 3. Matrix implementation of the bank of comb filters (left plot) and the weighting applied to each row of the matrix (right plot)

Investigating a range of possible weighting functions based on desired tempo output responses, the most consistent results were obtained from an empirically derived *perceptual weighting* using the Rayleigh distribution function:

$$L_w[l] = \frac{l}{b^2} e^{\frac{-l^2}{2b^2}} \tag{7}$$

This gave a skewed response (shown in the right hand plot of Figure 3) with a maximum at a lag of 50 samples (where $b = 50$ in equation (7)) corresponding to the *preferred* human tempo [8] of 100bpm.

Because the filterbank did not change throughout the operation of the algorithm, it could be created offline. Since the weighting was also time-invariant, we chose to incorporate it directly into the filterbank by multiplying each row of M by $L_w[l]$ rather than introduce unnecessary computation by weighting the ACF at each iteration.

4. BEAT ALIGNMENT AND PREDICTION

We wish to evaluate our beat-tracking system by synthesizing beats in time with the input audio. The result of the beat period estimation alone is insufficient information to successfully create a beat output. Equal importance must be given to the placement of the beats as well as their underlying rate [14], both of which are required to enable the prediction of future beats in the input signal.

4.1. Beat Alignment

A search for the best matching beat alignment over all possible shifts of the beat period (up to one whole period) is performed by cross-correlation of the detection function with the comb filter matched to the beat period. In order

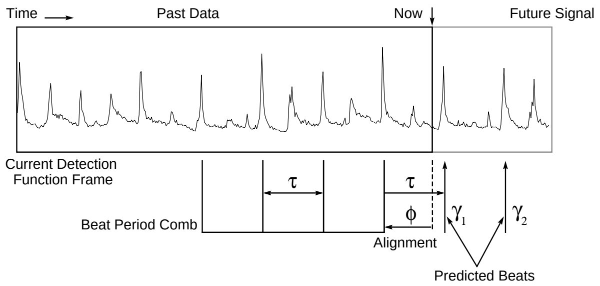

Figure 4. Beat alignment and prediction of a single analysis frame of the detection function. Beats are predicted only up the end of the subsequent step increment

to preserve the accuracy of the beat period, we need subsample accuracy in the alignment process. We upsample the detection function and the comb filter by a factor of 4, calculating an offset $\phi_{\max}$ with respect to a reference point in the current detection function frame:

$$4 \times \phi_{\max} = \arg\max_{\phi} \sum_{\phi=1}^{\Phi} \delta(n - 4k\tau + \phi) df_4[n] \qquad (8)$$

where $k = 1 : 4$ (to make four peaks in the comb filter), τ is the beat period, ϕ represents the shifts up to one period Φ and $df_4[n]$ is the upsampled detection function. To maintain causality and give the most salient alignment, the offset indicates the location of the *last* beat before the end of the current frame, see Figure 4.

4.2. Beat Prediction

For the beat tracking to be causal, no future audio data can be used in the analysis, therefore future beats must be predicted from past data. Our system performs beat prediction by triggering synthesized beats at intervals specified by the beat period from temporal locations defined by the alignment process. All beats are aligned with reference to the end of the current analysis frame, and are predicted only to the end of the next step increment. Any predictions beyond this point should be superseded in accuracy by the analysis of the following frame. A graphical representation of the beat prediction process is given in Figure 4, with the beat predictions calculated using the formula:

$$\hat{\gamma}_m = (t_i - \phi_i) + m\tau_i \qquad (9)$$

where $\hat{\gamma}_m$ is location of the m^{th} predicted beat in the subsequent analysis frame $(i + 1)$, t_i is the time at the end of frame i, with ϕ_i, the offset and τ_i, the beat period of frame i. To compensate for the long window size (about 6 seconds), a full frame of white noise (to simulate background noise) is prepended to the detection function prior to runtime. This is to allow some *approximate* tempo analysis to occur before a full time frame has elapsed.

Musical Genre	Tempo Accuracy (%)	Beat Accuracy (%)	Alignment Loss (%)
Rock (6)	89	87	2
Funk (4)	89	87	2
Electronic (4)	90	79	11
Jazz (4)	86	84	2
Classical (3)	86	63	23
Pop (3)	87	83	4
Total (24)	88	81	7

Table 1. Summary of Tempo Tracking Performance. The number of songs used in each genre is shown in brackets

5. RESULTS

The performance of the tempo tracking system was evaluated by subjective comparison with beats tapped by a musician. By assuming *perfect* beat placement from the human tapped performance, a simple accuracy measure is derived based on the ratio of *good* predictions to the total number of predicted beats. The errors correspond to spurious beats (including those with no temporal relevance) and omitted beats. Data is also presented based on the beat period estimates, to highlight the accuracy in the tempo analysis stage, and demonstrate the discrepancy between the tempo analysis and beat prediction stages. The results are summarized in Table 1. The test set included commercial and non-commercial recordings (sampled at 44.1kHz) covering a range of musical genres each lasting between 30 seconds and 1 minute in length.

An immediate observation from the results shown in Table 1, is that, in every case, the beat accuracy is lower than the tempo accuracy. This implies that more errors are made in the alignment and prediction stages than simply those carried forward from incorrect beat period estimates. It appears that the algorithm is entirely reliant on a well structured detection function, and that the majority of errors were caused by deviations from consistent rhythmic structure, particularly areas of low musical activity such as song intros. Similarly the first few beats were often the least accurately predicted, due to beats being derived from a combination of the detection function and the pre-pended noise, where the past data naturally lacks structure. These results were not amended to discount the initial errors because a similar approach would be needed for a real-time implementation of the algorithm. We have not yet performed a comprehensive evaluation against other systems, but initial tests against a version of Scheirer's [14] implementation indicate we are obtaining competitive results.

6. COMPLICATED BEAT TRACKING PROBLEMS

The aim of this section was to test the robustness of the algorithm when applied to a number of difficult test cases, created to simulate the possible behaviour of an impro-

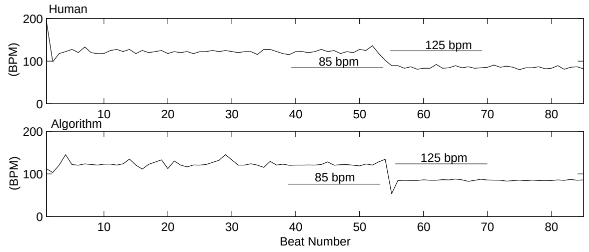

Figure 5. Comparison of human tracking (upper plot) with algorithm tracking (lower plot) for step tempo change

vised performance to which some automatic accompaniment might be desired. The test cases comprised a step tempo change, a ramp tempo change and an expressive piano performance. Subjective comparisons were made between the algorithm's performance and beats tapped by a musician. Plots are given showing tempo contours, derived from the inter-onset-intervals both for human tapped and algorithm predicted beats.

6.1. Step Tempo Change

To simulate a step tempo change, two songs of different fundamental tempo were cut together: a song with an electronic rhythm track (125 bpm) followed by a funk song (85 bpm). Figure 5 shows clearly when the transition between the two songs occurs. There is a noticeable delay of approximately 2 seconds before the system detects the tempo of the second segment. However outside of the transition the system performs with an accuracy equivalent to when the input signal has no abrupt tempo change.

6.2. Ramp Tempo Change

A ramp tempo change may be characterised by a continued acceleration or deceleration in the tempo of a piece of music. An example is the song "L.A. Woman" recorded by The Doors. Figure 6 shows the performance of the algorithm is not as consistent as the human tapped beats, which was noticeable when the beat tracks were compared audibly. The beat prediction stage assumes the tempo is constant across the length of the window, this means linear tempo progressions can only be approximated by *step-like* transitions. A listening test confirmed the predicted beats lagged behind the human tapped beats in the excerpt.

An interesting result occurred in this case, related to the *tempo octave problem* [11], exhibited by the sudden *jump* in tempo in the predicted beat track. Listening tests confirmed that no such discontinuity in the tempo existed. The apparent change in tempo was induced by the lag weighting applied to the filterbank (section 3.3), where *half* the beat period was found to be a better match to the input data than the actual tempo - indicating an upper limit for the system of around 170 bpm. We intend to investigate this effect further.

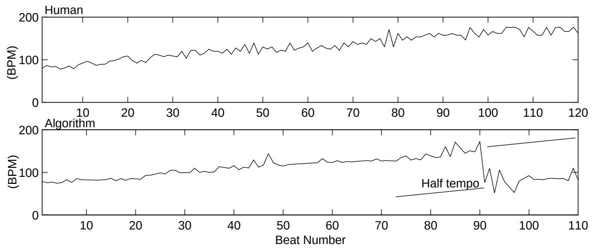

Figure 6. Comparison of human tracking (upper plot) with algorithm tracking (lower plot) for ramp tempo change

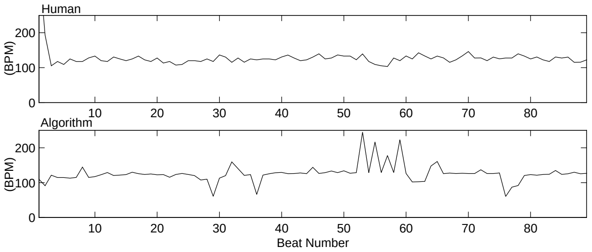

Figure 7. Comparison of human tracking (upper plot) with algorithm tracking (lower plot) for expressive piano music

6.3. Expressive Piano Performance

The final case study examines an excerpt of expressive solo piano (Mozart's Sonata K. 310). Due to the inconsistency in the tempo of the piece it is difficult to make a valid visual comparison of the tempo contours shown in Figure 7. The collection of spikes in the predicted beat plot are the result of a change in rhythmic structure, where the system switches between tracking at the underlying rate and twice the tempo. The approximate beat accuracy measure for the piece was found to be 75% which indicates moderate performance when compared to the test subjects in Table 1 with approximately constant tempo. The difficulty in accurate evaluation of this case in particular, highlights the need for more robust evaluation, a detailed discussion of evaluation procedures related to beat tracking is given in [9].

7. DISCUSSION

The algorithm's performance is based entirely on robust short term tempo analysis. Therefore the most common errors in beat period occur when the detection function, and hence the ACF are poorly defined. In the alignment stage, additional errors occur where the maximal cross-corelation is aligned to a strong accent (i.e. a large peak in the detection function) which is not related to the beat structure of the song. Due to the nature of the prediction process, errors in either beat period or alignment are repeated when more than a single beat is predicted per frame. Our analysis suggests that a time window of 6 seconds is enough prior information to correctly deduce the tempo, but that the alignment of tapped beats would be more accurately tracked by incorporating feedback into the system. Current work is underway to reduce alignment errors by using previously predicted beats to create an expectation of where subsequent alignments should occur. The result of which should also impose some notion of continuity to the system - an important factor which is not addressed in this implementation.

A number of further improvements might increase the performance of the tempo tracking system. Firstly a rhythmic confidence measure on the input ACF could be used to indicate the extent to which the data is well-structured and hence reliable. This could be applied to a more intelligent use of the two detection functions where, instead of simply combining them, a differential selection could be made based on the confidence of each function. It would also be interesting to examine perceptual studies of tempo tracking, to construct a more theoretically well-founded perceptual weighting for the filterbank as well as implementing alternative onset analysis stages, such as an envelope based approach [14] or energy flux [11]. Generally the system would benefit from a more robust evaluation stage including a direct comparison with other approaches to the problem.

8. CONCLUSIONS

We have introduced a causal approach to tempo tracking for musical audio signals. The system has been shown to approximate human tapped performance over a range of musical styles for signals with roughly constant tempo, with encouraging results for the more complicated tracking problems exhibiting tempo change.

Further research is underway to evaluate the system by comparison with alternative approaches and improve performance by incorporating high level musical knowledge into the system as well as developing a real-time implementation.

9. ACKNOWLEDGEMENTS

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

10. REFERENCES

[1] Allen P. E. and Dannenberg R. B. "Tracking Musical Beats in Real Time", *Proceedings of International Computer Music Conference*, pp. 140-143, 1990

[2] Bello J. P., Daudet L., Abdallah S., Duxbury C., Davies M. E. and Sandler M. B. "A Tutorial on Onset Detection in Music Signals", *IEEE Transactions on Speech and Audio Processing - accepted for publication*, 2004

[3] Bello J.P., Duxbury C., Davies M.E. and Sandler M.B., "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, June 2004

[4] Brossier P., Bello J. P. and Plumbley M. D. "Real-time Temporal Segmentation of Note Objects in Music Signals", *Proceedings of International Computer Music Conference*, 2004

[5] Brown J. C. "Determination of the Meter of Musical Scores by Autocorrelation" *Journal of the Acoustical Society of America*, Vol 94, No. 4, 1993

[6] Dixon S. "Automatic Extraction of Tempo and Beat from Expressive Performances" *Journal of New Music Research*, Vol 30, No. 1, March 2001

[7] Duxbury C. "Signal Models for Polyphonic Music" *Ph.D. Thesis (submitted), Department of Electronic Engineering, Queen Mary, University of London*, 2004

[8] Fraisse P. "Rhythm and Tempo", In D. Deutsch, editor, *The Psychology of Music*, Academic Press, 1982

[9] Goto M. and Muraoka Y. "Issues in Evaluating Beat Tracking Systems" *Working Notes of IJCAI-97 Workshop on Issues in AI and Music - Evaluation and Assessment*, 1997

[10] Goto M. "An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds" *Journal of New Music Research*, Vol 30, No. 2, July 2001

[11] Laroche J. "Efficient Tempo and Beat Tracking in Audio Recordings" *Journal of the Audio Engineering Society*, Vol 51, No. 4, 2003

[12] Masri P. and Bateman A. "Improved Modelling of Attack Transients in Music Analysis-Resynthesis" *Proceedings of International Computer Music Conference*, pp. 100-103, 1996

[13] Raphael C. "Automated Rhythm Transcription" *Proceedings of International Symposium on Music Information Retrieval*, pp. 99-107, 2001

[14] Scheirer E. D. "Tempo and Beat Analysis of Acoustic Musical Signals " *Journal of the Acoustical Society of America*, Vol 103, No. 1, 1998