Tempo and Beat Estimation of Musical Signals

Conference Paper · October 2004

3 authors:

Miguel A. Alonso
Center for Scientific Research and Higher Education at Ensenada

Bertrand David
Institut Mines-Télécom

Gaël Richard
Télécom Paris

TEMPO AND BEAT ESTIMATION OF MUSICAL SIGNALS

Miguel Alonso, Bertrand David , Gaël Richard

ENST-GET, Département TSI
46, rue Barrault, Paris
75634 cedex 13, France
{malonso,bedavid,grichard}@tsi.enst.fr

ABSTRACT

Tempo estimation is fundamental in automatic music processing and in many multimedia applications. This paper presents an automatic tempo tracking system that processes audio recordings and determines the beats per minute and temporal beat location. The concept of spectral energy flux is defined and leads to an efficient note onset detector. The algorithm involves three stages: a front-end analysis that efficiently extracts onsets, a periodicity detection block and the temporal estimation of beat locations. The performance of the proposed method is evaluated using a large database of 489 excerpts from several musical genres. The global recognition rate is 89.7 %. Results are discussed and compared to other tempo estimation systems.

Keywords: *beat, tempo, onset detection.*

1. INTRODUCTION

It is very difficult to understand western music without perceiving beats, since a beat is a fundamental unit of the temporal structure of music [4]. For this reason, automatic beat tracking, or tempo tracking, is an essential task for many applications such as musical analysis, automatic rhythm alignment of multiple musical instruments, cut and paste operations in audio editing, beat driven special effects. Although it might appear simple at first, tempo tracking has proved to be a difficult task when dealing with a broad variety of musical genres as shown by the large number of publications on this subject appeared during the last years [2, 5, 6, 8, 9, 10, 12].

Earlier tempo tracking approaches focused on MIDI or other symbolic formats, where note onsets are already available to the estimation algorithm. More recent approaches directly deal with ordinary CD audio recordings. The system that we present in this paper lies into this category.

For musical genres with a straightforward rhythm such as rap, rock, reggae and others where a strong percussive strike drives the rhythm, current beat trackers indicate high performance as pointed out by [5, 9, 12]. However, the robustness of the beat tracking systems is often much less guaranteed when dealing with classical music because of the weakness of the techniques employed in attack detection and tempo variations inherent to that kind of music.

In the present article, we describe an algorithm to estimate the tempo of a piece of music (in beats per minute or bpm) and identify the temporal locations when it occurs. Like most of the systems available in the literature, this algorithm relies on a classical scheme: a front-end processor extracts the onset locations from a time-frequency or subband analysis of the signal, traditionally using a filter bank [1, 7, 10, 12] or using the discrete Fourier transform [3, 5, 6, 8, 9]. Then, a periodicity estimation algorithm finds the rate at which these events occur. A large variety of methods has been used for this purpose, for example, a bank of oscillators which resonate at integer multiples of their characteristic frequency [6, 9, 12], pitch detection methods [1, 10], histograms of the inter-onset intervals [2, 13], probabilistic approaches such as Gaussian mixture model to express the likelihood of the onset locations [8].

In this paper, following Laroche's approach [9], we define the quantity so-called spectral energy flux as the derivative of the signal frequency content with respect to time. Although this principle has been previously used [3, 6, 8, 9], a significant improvement has been obtained by using an optimal filter to approximate this derivative.

We exploit this approach to obtain a high performance onset detector and integrate it into a tempo tracking algorithm. We demonstrate the usefulness of this approach by validating the proposed system using a large manually annotated data base that contains excerpts from rock, latin, pop, soul, classical, rap/hip-hop and others. The paper is organized as follows: Section 2 provides a detailed description of the three main stages that compose the system. In Section 3, test results are provided and compared to other methods. The system parameters used during the

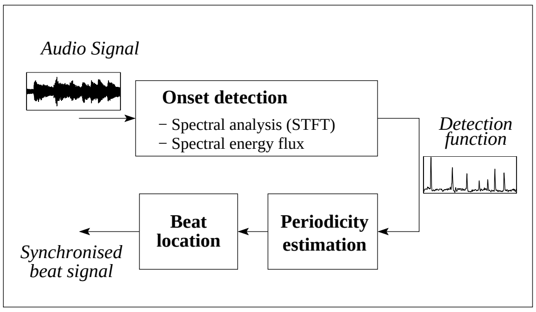

Figure 1. Architecture of the beat tracking algorithm.

validation procedure are provided as well as comments about the issues of the algorithm. Finally, Section 4 summarizes the achievements of the presented algorithm and discusses possible directions for further improvements.

2. DESCRIPTION OF THE ALGORITHM

In this paper, it is assumed that the tempo of the audio signal is constant over the duration of the analysis window and that it eventually evolves slowly from one to the other. In addition, we suppose that the tempo lies between 60 and 200 BPM, without loss of generality since any other value can be mapped into this range. The algorithm proposed is composed of three major steps (see figure 1):

- *onset detection*: it consists in computing a detection function based on the spectral energy flux of the input audio signal;
- *periodicity estimation* : the periodicity of the detection function is estimated using pitch detection techniques ;
- *beat location estimation* : the position of the corresponding beats is obtained from the cross-correlation between the detection function and an artificial pulse-train.

2.1. Onset detection

The aim of onset detection consists in extracting a detection function that will indicate the location of the most salient features of the audio signal such as note changes, harmonic changes and percussive events. As a matter of fact, these events are particularly important in the beat perception process.

Note onsets are easily masked in the overall signal energy by continuous tones of higher amplitude [9], while they are more likely detected after separating them in frequency channels. We propose to follow a frequency domain approach [3, 5, 6, 8, 9] as it proves to outperform time-domain methods based on direct processing of the temporal waveform as a whole.

2.1.1. *Spectral analysis and spectral energy flux*

The input audio signal is analyzed using a decimated version of the short-time Fourier transform (STFT), i.e., short signal segments are extracted at regular time intervals, multiplied by an analysis window and transformed into the frequency domain by means of a Fourier transform. This leads to

$$\tilde{X}(f,m) = \sum_{n=0}^{N-1} w(n)x(n+mM)e^{-j2\pi fn} \tag{1}$$

where $x(n)$ denotes the audio signal, $w(n)$ the finite analysis window of size N in samples, M the hop size in samples, m the frame index and f the frequency.

Motivated by the work of Laroche [9], we define the spectral energy flux $E(f,k)$ as an approximation to the derivative of the signal frequency content with respect to time

$$E(f,k) = \sum_m h(m-k)\,G(f,m) \tag{2}$$

where $h(m)$ approximates a differentiator filter with:

$$H(e^{j2\pi f}) \simeq j2\pi f \tag{3}$$

and the transformation

$$G(f,m) = \mathcal{F}\{|\tilde{X}(f,m)|\} \tag{4}$$

is obtained via a two step process: *a low-pass filtering* of $|\tilde{X}(f,m)|$ to perform energy integration in a way similar to that in the auditory system, emphasizing the most recent inputs, but masking rapid modulations [14] and *a non-linear compression*. For example, in [9] Laroche proposes $h(m)$ as a first order differentiator filter ($h=[1;-1]$), no low-pass filtering is applied and the non-linear compression function is $G(f,n) = \text{arcsinh}(|\tilde{X}(f,m)|)$. In [6] Klapuri uses the same first order differentiator filter, but for the transformation, he performs the low-pass filtering after applying a logarithmic compression function.

In the present work we propose $h(m)$ to be a FIR filter differentiator. Such a filter is designed by a Remez optimisation procedure which leads to the best approximation to Eq. (3) in the minimax sense [11]. This new approach, compared to the first order difference used in [6, 8, 9] highly improves the extraction of musical meaningful features such as percussive attacks and chord changes. In addition, $G(f,k)$ is obtained via low-pass filtering with a second half of a Hanning window followed by a logarithmic compression function as suggested by Klapuri [7], since the logarithmic difference function gives the amount of change in a signal in relation to its absolute level. This is a psycho-acoustic relevant measure since the perceived signal amplitude is in relation to its level, the same amount of increase being more prominent in a quite signal [7].

During the system development, several orders for the differentiator filter $h(m)$ were tested. We found that using an order 8 filter was the best tradeoff between complexity and efficiency. In practice, the algorithm uses an N

point FFT to evaluate the STFT, thus the frequency channels 1 to $\frac{N}{2}$ of the signal's time–frequency representation are filtered using $h(m)$ to obtain the spectral energy flux. Then, all the positive contributions of these channels are summed to produce a temporal waveform $v(k)$ that exhibits sharp maxima at transients and note onsets, i.e., those instants where the energy flux is large.

Beat tends to occur at note onsets, so we must first distinguish the "true beat" peaks from the spurious ones in $v(k)$ to obtain a proper detection function $p(k)$. In addition, we work under the supposition that these unwanted peaks are much smaller in amplitude compared to the note attack peaks. Thus, a peak-picking algorithm that selects peaks above a dynamic threshold calculated with the help of a median filter is a simple and efficient solution to this problem. The median filter is a nonlinear technique that computes the pointwise median inside a window of length $2i+1$ formed by a subset of $v(k)$, thus the median threshold curve is given by the expression:

$$\theta(k) = \mathcal{C} \cdot \text{median}(g_k) \tag{5}$$

where $g_k = \{v_{k-i}, \ldots, v_k, \ldots, v_{k+i}\}$ and $\mathcal{C}$ is a predefined scaling factor to artificially rise the threshold curve slightly above the steady state level of the signal. To ensure accurate detection, the length of the median filter must be longer than the average width of the peaks of the detection function. In practice, we set the median filter length to 200 ms. Then, we obtain the signal $\hat{p}(k) = v(k) - \theta(k)$, which is half-wave rectified to produce the detection function $p(k)$:

$$p(k) = \begin{cases} \hat{p}(k) & \text{if } \hat{p}(k) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

In our tests, the onset detector described above has proved to be a robust scheme that provides good results for a wide range of musical instruments and attacks at a relatively low computational cost. For example, Figure 2-a shows the time waveform of a piano recording containing seven attacks. These attacks can be easily observed in the signal's spectrogram in Figure 2-b. The physical interpretation of Figure 2-c can be seen as the rate at which the frequency-content energy of the audio signal varies at a given time instant, i.e., the spectral energy flux. In this example, seven vertical stripes represent the reinforcement of the energy variation, clearly indicating the location of the attacks (the position of the spectrogram edges). When all the positive energy variations are summed in the frequency domain and thresholded, we obtain the detection function $p(k)$ shown in Figure 2-d. An example of an instrument with smooth attacks, a violin, is shown in Figure 3. Large energy variations in the frequency content of the audio signal can still be observed as vertical stripes in Figure 3-c. After summing the positive contributions, six of the seven attacks are properly detected as shown by the corresponding largest peaks in Figure 3-d.

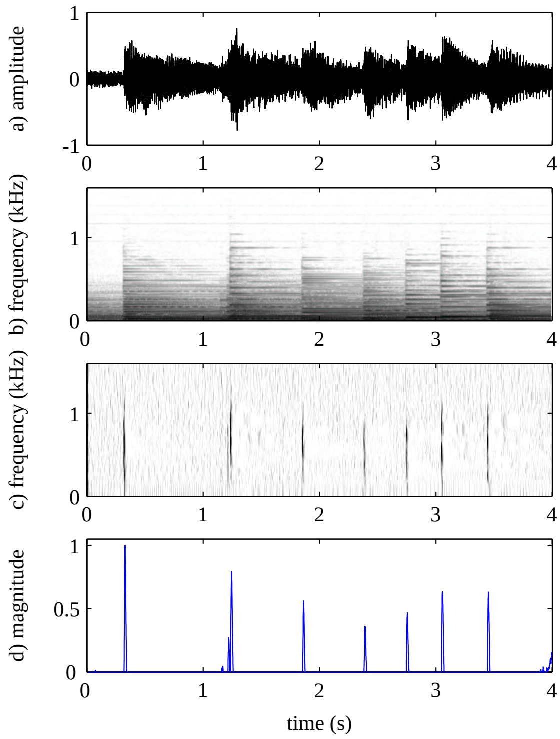

Figure 2. From top to bottom: time waveform of a piano signal, its spectrogram, its spectral energy flux and the detection function $p(k)$.

2.2. Periodicity estimation

The detection function $p(k)$ at the output of the onset detection stage can be seen as a quasi-periodic and noisy pulse-train that exhibits large peaks at note attacks. The next step is to estimate the tempo of the audio signal, i.e., the periodicity of the note onset pulses. Two methods from traditional pitch determination techniques are employed: the spectral product and the autocorrelation function. These techniques have already been used for this purpose in [1].

2.2.1. Spectral product

The spectral product principle assumes that the power spectrum of the signal is formed from strong harmonics located at integer multiples of the signal's fundamental frequency. To find this frequency, the power spectrum is compressed by a factor m, then the obtained spectra are multiplied, leading to a reinforced fundamental frequency. For a normalized frequency, this is given by:

$$S(e^{j2\pi f}) = \prod_{m=1}^{M} |P(e^{j2\pi m f})| \quad \text{for } f < \frac{1}{2M} \tag{7}$$

where $P(e^{j2\pi f})$ is the discrete Fourier transform of $p(k)$. Then, the estimated tempo $\mathbb{T}$ is easily obtained by picking out the frequency index corresponding to the largest peak of $S(e^{j2\pi f})$. The tempo is constrained to fall in the range $60 < \mathbb{T} < 200$ BPM.

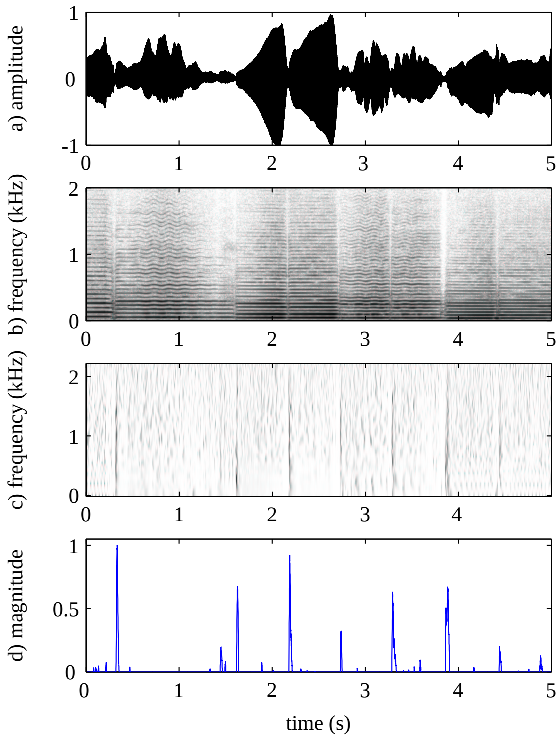

Figure 3. From top to bottom: time waveform of a violin signal, its spectrogram, its spectral energy flux and the detection function $p(k)$.

2.2.2. Autocorrelation function

This is a classical method in periodicity estimation. The non-normalized deterministic autocorrelation function of $p(k)$ is calculated as follows:

$$r(\tau) = \sum_k p(k+\tau)p(k) \tag{8}$$

Again, we suppose that $60 < \mathbb{T} < 200$ BPM. Hence, during the calculation of the autocorrelation, only the values of $r(\tau)$ corresponding to the range of 300 ms to 1 s are calculated. To find the estimated tempo $\mathbb{T}$, the lag of the three largest peaks of $r(\tau)$ are analyzed and a multiplicity relationship between them is searched. In the case that no relation is found, the lag of the largest peak is taken as the beat period.

2.3. Beat location

To find the beat location, we use a method based on the comb filter idea that resembles previous work carried out by [6, 9, 12]. We create an artificial pulse-train $q(t)$ of tempo $\mathbb{T}$ previously calculated as explained in Section 2.2 and cross-correlate it with $p(k)$. This operation has a low computational cost, since the correlation is evaluated only at the indices corresponding to the maxima of $p(k)$. Then, we call t_0 the time index where this cross-correlation is maximal and we consider it as the starting location of the beat. For the second and succesive beats in the analysis window, a beat period $\mathcal{T}$ is added to the previous beat location, i.e., $t_i = \lfloor t_{i-1} + \mathcal{T} \rfloor$ and a corresponding peak in $p(k)$ is searched within the area $t_i \pm \Delta$. If no peak is found, the beat is placed in its expected position t_i. When the last beat of the window occurs, its location is stored in order to assure the continuity with the first beat of the new analysis window. Where the tempo of the new analysis window differs by more than 10 % from the previous tempo, a new beat phase is estimated. The peaks are searched using the new beat period without referencing the previous beat phase.

Genre	Pieces	Percentage
classical	137	28.0 %
jazz	79	16.2 %
latin	37	7.6 %
pop	40	8.2 %
rock	44	9.0 %
reggae	30	6.1 %
soul	24	4.9 %
rap, hip-hop	20	4.1 %
techno	23	4.7 %
other	55	11.2 %
total	**489**	**100 %**

Table 1. Genre distribution of the test database.

3. PERFORMANCE ANALYSIS

3.1. Database, annotation and evaluation protocole

The proposed algorithm was evaluated using a corpus of 489 musical excerpts taken from commercial CD recordings. These pieces were selected to cover as many characteristics as possible: various tempi in the 50 to 200 BPM range, a wide variety of instruments, dynamic range, studio/live recordings, old/recent recordings, with/without vocals, male/female vocals and with/without percussions. They were also selected to represent a wide diversity of musical genres as shown in Table 1.

From each of the selected recordings, an excerpt of 20 seconds having a relatively constant tempo, was extracted and converted to a monophonic signal sampled at 16 kHz. The procedure for manually estimating the tempo of each musical piece is the following:

- the musician listens to a musical excerpt using headphones (if required, several times in a row to be accustomed to the tempo),
- while listening, he/she taps the tempo,
- the tapping signal is recorded and the tempo is extracted from it.

As pointed out by Goto in [4], the beat is a perceptual concept that people feel in music, so it is generally difficult to define the "correct beat" in an objective way. People have a tendency to tap at different metric levels. For

Method	Recognition rate
Paulus [10]	56.3 %
Scheirer [12]	67.4 %
SP .	63.2 %
AC .	73.6 %
SP using SEF.	84.0 %
AC using SEF	89.7 %

Table 2. Tempo estimation performances. SEF stands for spectral energy flux, SP for spectral product and AC for autocorrelation.

Method Genre	PLS %	SCR %	SP %	AC %	SP-SEF %	AC-SEF %
classical	46.0	46.2	48.2	70.8	71.5	82.4
jazz	57.0	70.9	62.0	69.8	78.4	86.0
latin	70.3	81.1	62.1	70.3	91.8	94.5
pop	57.5	70.0	75.0	85.7	92.5	92.5
rock	40.9	84.1	61.3	84.4	81.8	88.6
reggae	76.7	86.7	86.6	76.9	96.6	100
soul	50.0	87.5	70.8	76.7	100	100
rap	75.0	85.0	75.0	56.5	100	100
techno	69.6	56.3	65.2	95.0	95.6	100
other	61.8	69.1	74.5	66.7	89.0	90.9

Table 3. Tempo estimation performances by musical genre. PLS stands for Paulus [10], SCR for Scheirer [12].

example, in a piece that has a 4/4 time signature, it is correct to tap every quarter-note or every half-note. In general, a ”ground truth” tempo cannot be established unless the musical score of the piece is available. This is a very common problem when humans tap along with the music, i.e., to tap twice as fast or twice as slow the ”true” tempo. Whenever this case ocurred during the database annotation, the slower tempo was taken as reference $\mathbb{T}_R$. In a similar way to humans, automatic tempo estimation methods also make this doubling or halving of the ”true” tempo. Thus, for evaluation purposes the tempo estimation $\mathbb{T}$ provided by the algorithm is labeled as correct if there is a less than 5% disagreement from the manually annotated tempo used as reference $\mathbb{T}_R$ under the principle $0.95\alpha\mathbb{T} < \mathbb{T}_R < 1.05\alpha\mathbb{T}$ with $\alpha \in \{\frac{1}{2}, 1, 2\}$.

3.2. Results

During the evaluation, the algorithm parameters were set as follows. The length of the analysis window for tempo estimation was set to four seconds, with an overlapping factor of 50%. Smaller window size values reduce the algorithm performance. For the spectral energy flux calculation, the length of the analysis window used in the computation of the STFT was 64 samples (4 ms) with an overlapping factor of 50% and a 128 point FFT, thus the detection function $v(k)$ could be seen as signal sampled at 500 Hz. As mentioned, the order of the differentiator FIR filter was set to $L = 8$. In the beat location stage, the median filter i was set to 25 samples, C was set to 2, and for the peak location Δ was set to 10 % of the beat period.

To have a better idea of the performance of our algorithm, we decided to compare it with our own implementation of the algorithms proposed by Paulus [10] and Scheirer [12]. We also compared it with our previous work in tempo estimation [1]. In this case, the main difference between the previous and the current system lies on the onset extraction stage. Table 2 summarizes the overall recognition rate for the evaluated systems. In this table, SP stands for spectral product, AC for autocorrelation and SEF for spectral energy flux.

In more details, the performance of these methods by musical genre are presented in Table 3. In this table, PLS stands for Paulus, SCR for Scheirer. As expected, results indicate that classical music is the most difficult genre. Nevertheless, the proposed algorithm displayed promising results. For the other genres, it shows good performance, particularly for music with a straightforward rhythm.

Several authors have pointed out the difficulty in evaluating beat tracking systems [4, 6, 9] due to the subjective interpretation of the beat and the inexistence of a consensual data base of beat-labeled audio tracks. In our case, the beat location evaluation was done at a subjective level, that is, artificial ”sound clicks” were superimposed on the tested signal at the calculated beat locations and tempo.

During the validation procedure, we note that the proposed algorithm produces erroneous results under the following circumstances:

- when dealing with signals having a stealthily or long fading-in attacks, the hypothesis that supurious peaks are smaller than attack peaks does not hold any more, leading to false onset detections;
- the spectral energy flux follows the principle that stable spectra regions are followed by transition regions. When many instruments play simultaneously, as in an orchestra, their ’spectral mixture’ lacks stable regions, leading to false onset detections;
- when the tempo varies too quickly in short time segments or if there are large beat gaps in the signal, the periocity estimation stage cannot keep up with the changes.

The reader is welcome to listen to the sound examples available at `www.tsi.enst.fr/~malonso/ismir04`.

4. CONCLUSIONS

In this paper we have presented an efficient beat tracking algorithm that processes audio recordings. We have also defined the concept of spectral energy flux and used it to derive a new and effective onset detector based on the STFT, an efficient differentiator filter and dynamic thresholding using a median filter. This onset detector displays high performance for a large range of audio signals. In addition, the proposed tempo tracking system is straightforward to implement and has a relatively low computational cost. The performance of the algorithm presented was evaluated on a large database containing 489 musical excerpts from several musical genres. The results are encouraging since the global success rate for tempo estimation was 89.7%. The method presented works offline. A real-time implementation is considered, but currently there are various issues to be resolved such as the block-wise processing that requires access to future signal samples and the non-causality of the thesholding filter. Future work should explore other periodicity estimation techniques and an analysis of the residual part after a harmonic/noise decomposition.

5. REFERENCES

[1] Alonso M., David B. and Richard G., ”A Study of Tempo Tracking Algorithms from Polyphonic Music Signals”, *Proceedings of the 4th. COST 276 Workshop*, Bordeaux, France. March 2003.

[2] Dixon, M. ”Automatic Extraction of Tempo and Beat from Expressive Performances”, *Journal of New Music Research*,vol. 30 No. 1, pp. 39–58 2001.

[3] Duxbury C., Sandler M. and Davies M., ”A Hybrid Approach to Musical Note Onset Detection”, *Proceedings of the 5th. Int. Conf. on Digital Audio Effects (DAFx)*, pp. 33–38, Hamburg Germany, September 2002.

[4] Goto M. and Muraoka Y. ”Issues in Evaluating Beat Tracking Systems”, *Working Notes of the IJCAI-97 Workshop on Issues in AI and Music*, August 1997.

[5] 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, pp. 159–171, 2001.

[6] Klapuri, A. ”Musical meter estimation and music transcription”, *Cambridge Music Processing Colloquium*, Cambridge University, UK, March 2003.

[7] Klapuri, A. ”Sound Onset Detection by Applying Psychoacoustic Knowledge”, *Proceedings IEEE Int. Conf. Acoustics Speech and Sig. Proc. (ICASSP)*, pp. 3089–3092, Phoenix AR, USA March 1999.

[8] Laroche, J. ”Estimating, Tempo, Swing and Beat Locations in Audio Recordings”, *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 135–138, New Paltz, NY, USA October 2001.

[9] Laroche, J. ”Efficient Tempo and Beat Tracking in Audio Recordings”, *J. Audio. Eng. Soc.*, vol. 51, No. 4, pp. 226–233, April 2003.

[10] Paulus J. and Klapuri A., ”Measuring the Similarity of Rythmic Patterns”, *Proceedings of the International Symposium on Music Information Retrieval*, Paris, France, 2002.

[11] Proakis, J. G. and Manolakis, D., *Digital Signal Processing: Principles, Algorithms and Applications*. Prentice Hall, New York, 1995.

[12] Scheirer, E. D., ”Tempo and Beat Analysis of Acoustic Music Signals”, *J. Acoust. Soc. Am.*, vol. 104, pp. 588–601, January 1998.

[13] Seppa anen, J. ”Tatum grid analysis of musical signals”, *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA October 2001.

[14] Todd, N. P. McA., ”The Auditory ’Primal Sketch’: A Multiscale model of rhythmic grouping”, *Journal of New Music Research*, vol. 23, No. 1, pp. 25 – 70, 1994.