BEAT TRACKING USING GROUP DELAY BASED ONSET DETECTION

Andre Holzapfel and Yannis Stylianou
Institute of Computer Science, FORTH, Greece,
and Multimedia Informatics Lab, Computer Science Department, University of Crete
{hannover, yannis}@csd.uoc.gr

ABSTRACT

This paper introduces a novel approach to estimate onsets in musical signals based on the phase spectrum and specifically using the average of the group delay function. A frame-by-frame analysis of a music signal provides the evolution of group delay over time, referred to as phase slope function. Onsets are then detected simply by locating the positive zero-crossings of the phase slope function. The proposed approach is compared to an amplitude-based onset detection approach in the framework of a state-of-the-art system for beat tracking. On a data set of music with less percussive content, the beat tracking accuracy achieved by the system is improved by 82% when the suggested phase-based onset detection approach is used instead of the amplitude-based approach, while on a set of music with stronger percussive characteristics both onset detection approaches provide comparable results of accuracy.

1 INTRODUCTION

The task of estimating the times at which a human would tap his foot to a musical sound is known as beat tracking [1]. All state-of-the-art approaches ([1, 2, 3, 4]) for this task first conduct an onset detection. The output of the onset detection is a signal with a lower time resolution than the input signal, which has peaks at the time instances where a musical instrument in the input started playing a note. Usually, this onset signal is derived from the amplitude of the signal, as in [1, 2, 3]. Only in [4], phase information is considered, by computing the phase deviation between neighboring analysis frames. Several approaches of computing onsets from musical signals are compared in [5], resulting in the conclusion that in general the moments of big positive change in the amplitude spectrum of the signal provide the most preferable estimators for the onsets. This is because using information contained in the complex spectrum or in the phase changes might lead to similar onset estimations, but using only the amplitude spectrum is preferred for computational reasons [5]. A similar conclusion can be drawn from the results of the MIREX 2007 Audio Onset Detection contest [1] , where most systems use only the amplitude information. Considering the results on complex music mixtures, which are the signals of interest in beat tracking, the usage of phase deviation by some systems does not improve the onset estimation accuracy [6].

As can be seen in the results depicted in [7] (Table II), the state-of-the-art approaches for beat tracking decrease significantly in accuracy, when applied to folk music. These music signals contain weaker percussive content than music of rock or disco styles. This problem is of particular importance when dealing with traditional dances as well, as they are often played using string or wind instruments only [8]. Based on the results obtained in [7], it is necessary to improve beat tracking on music with little percussive content. While a decrease in the case of jazz and classical music can partly be attributed to the complex rhythmic structure, rhythmic structure of folk music is simpler, and thus the decrease in this forms of music may be attributed solely to the problem of detecting onsets.

In [9], the negative derivative of the unwrapped phase, *i.e.* the group delay, is used to determine instants of significant excitation in speech signals. This approach has been further developed and used for the detection of clicks of marine mammals in [10]. There, it has been shown that pulses can be reliably detected by using group delay, even when the pulses have been filtered by a minimum phase system. Motivated by these works, we suggest the use of the group delay function for onset estimation in musical signals, and then use this estimation as input to a beat tracker based on the state-of-the-art system presented in [2]. The goal of this paper is to provide an improved beat tracking performance on musical signals with simple rhythmic structure and little or no percussive content. The proposed approach to consider phase information is novel in Music Information Retrieval, as previous approaches computed a time derivative of phase [5], while the suggested approach makes use of group delay, which is a derivative of phase over *frequency*. The group delay function is computed in frame-by-frame analysis and its average is computed for each frame. This results in time-domain signal referred to as phase slope function [10]. Onsets are then simply estimated by detecting the positive zero-crossings of the phase slope function. Therefore, the suggested approach does not require the use of time-dependent (adaptive) energy-based thresholds as the ampli-

[1] http://www.music-ir.org/mirex/2007/index.php/Audio_Onset_Detection

tude and phase based approaches for detecting the onsets [5].

In Section 2, the characteristics of the group delay function for minimum phase signals are shortly reviewed to support our motivation. In Section 3 we present a method to compute an onset signal for music based on the phase slope function and how this information has been incorporated into the state-of-the-art system suggested by Klapuri et al. [2] for beat tracking. Section 4 shows results of the proposed approach on artificial and music signals comparing the suggested phase-based approach with a widely used amplitude-based approach. Section 5 concludes the paper and discusses future work.

2 BASIS FOR THE PROPOSED METHOD

Consider a delayed unit sample sequence $x[n] = \delta[n - n_0]$ and its Fourier Transform $X(\omega) = e^{-j\omega n_0}$. The group delay is defined as:

$$\tau(\omega) = -\frac{d\phi(\omega)}{d\omega} \tag{1}$$

so the group delay for the delayed unit sample sequence is $\tau(\omega) = n_0 \ \ \forall\omega$, since the phase spectrum of the signal is $\phi(\omega) = -\omega n_0$. The average over ω of $\tau(\omega)$ provides n_0 which corresponds to the negative of the slope of the phase spectrum for this specific signal and to the delay of the unit sample sequence. An example of a delayed unit sample sequence with $n_0 = 200$ samples as well as the associated group delay function are depicted in Fig.1(a) and (b), respectively. In the above example the Fourier Transform has been computed considering the center of analysis window to be at $n = 0$. When the window center is moved to the right (closer to the instant $n = n_0$), the slope of the phase spectrum is increased (the average of the group delay function is decreased) reflecting the distance between the center of the analysis window and the position of the impulse. When the center of the analysis window is at $n = n_0$ then the slope is zero. Continuing moving the analysis window to the right the slope will start to increase (while the average of the group delay will decreased). In this way, the slope of the phase spectrum is a function of n. Note that the location of the zero-crossing of this function will provide the position of the non-zero value of the unit sample sequence independently of the amplitude value of the impulse.

In general, the average value of the group delay is determined by the distance between the center of the analysis window and the delay of the unit sample sequence, even when it has been filtered by a minimum phase system [9]. The group delay function will still provide information about this delay value as well information about the poles of the minimum phase system. In Fig. 1(c),(d) the output of the minimum phase signal and the associated group delay are depicted. The slope function will have a similar behavior to this described earlier for the unit sample sequence.

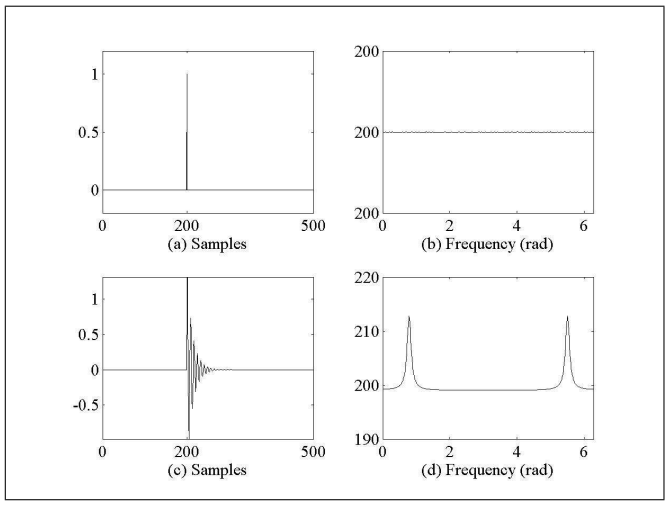

Figure 1. (a) A delayed by 200 samples unit sample sequence. (b) The group delay function of the signal in (a). (c) A minimum phase signal with an oscillation at $\pi/4$. (d) The group delay function of the signal in (c).

In Figure 2, the phase slope of a periodic sequence of minimum phase signals is depicted. The dash-dotted line depicts a phase slope resulting from a window shorter than the period of the signal, the dashed line results from an analysis using a longer window. The phase slope values have been assigned to the middle of the analysis window, the analysis step size was set to one sample. It can be seen that even in the case of low signal amplitude, the positive zero crossing coincides in each case with the beginning of the minimum phase signal. As a musical instrument could be considered as a causal and stable filter, that is driven by a minimum phase excitation, it can be assumed that an onset estimation using the positive zero crossings of the phase slope is a valid approach. To avoid the problems of unwrapping the phase spectrum of the signal for the computation of group delay, the slope of the phase function can be computed as [11]:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \tag{2}$$

where

$$\begin{aligned} X(\omega) &= X_R(\omega) + jX_I(\omega) \\ Y(\omega) &= Y_R(\omega) + jY_I(\omega) \end{aligned}$$

are the Fourier Transforms of $x[n]$ and $nx[n]$, respectively. The phase slope is then computed as the negative of the average of the group delay function.

3 METHOD FOR BEAT TRACKING

3.1 Onset detection using group delay

The onset detection using group delay follows the concept explained in Section 2. The parameters of the onset detector

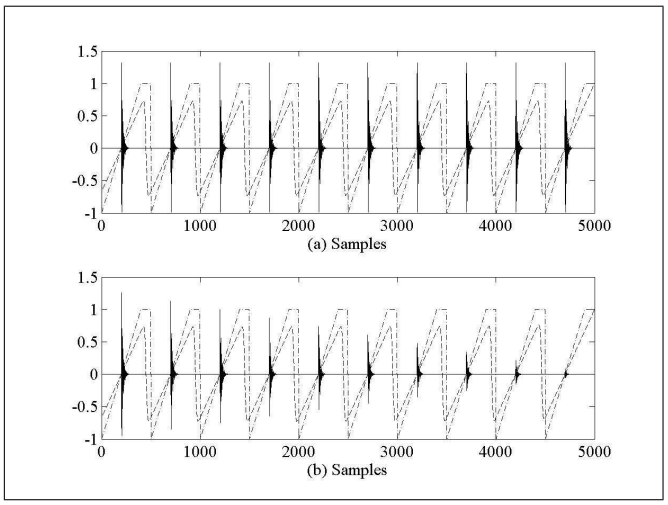

Figure 2. (a) A sequence of impulses of constant amplitude and the associated phase slope function using long (dashed line) and short (dash-dotted line) window (b) A sequence of impulses with linearly time varying amplitudes and the associated phase slope function using long (dashed line) and short (dash-dotted line) window.

have been evaluated on two development data sets: The first data set, referred to as D1, consists of periodic artificial signals like those depicted in Figure 2, with periods from 0.3s to 1s, which is related to the typical range for the tempo of musical pieces (60bpm-200bpm). This data set is also useful in evaluating the robustness of the suggested approach against additive noise. For this purpose, a Transient to Noise Ratio (TNR) is defined in the same way as the usual Signal to Noise Ratio (SNR):

$$TNR = 10 \log_{10} \frac{\frac{1}{L} \sum_{n=0}^{L-1} x^2(n)}{\sigma_u^2} \qquad (3)$$

where x denotes a signal of length L and σ_u^2 denotes the variance of the noise. The artificial signals have been mixed with white noise at different TNR. For each artificial signal a corresponding text file has been created containing the onset times of the impulses. The second development set, referred to as D2, is a data set of 28 song excerpts of 30 seconds length. This data set has been compiled and beat annotated by the authors. From these annotations, a unit sample sequence, $\mathbf{a}[n]$, may be obtained with pulses located at the annotated onset or beat time instance. In the same way, a unit sample sequence $\mathbf{y}[n]$ may be generated from the estimated onset times. The quality of an onset estimation was judged based on the function used in the MIREX 2006 Audio Beat Tracking contest [2] :

$$P_M = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{N_P} \sum_{m=-W}^{W} \sum_{n=1}^{N} \mathbf{y}[n] \mathbf{a}_s[n-m] \qquad (4)$$

[2] http://www.music-ir.org/mirex2006/index.php/Audio_Beat_Tracking

where N is the length of the two pulse trains in samples, S is the number of different annotations per sound sample (equal to one for the development data), N_P is the maximum number of impulses in the two pulse trains, $N_P = \max(\sum \mathbf{y}[n], \sum \mathbf{a}_s[n])$, and W is equal to 20% of the average distance between the impulses in $\mathbf{a}_s[n]$ in samples. This function represents an estimator, of how much two pulse trains are correlated, accepting some inaccuracy regarding the placement of the onset estimation impulses. Note that for the development set containing music signals (*i.e.*, D2), no onset annotations exist but beat annotations.

The most crucial parameter in the click or onset detection using phase slope is the length of the analysis window. As it is indicated in [10], a large window is appropriate for detecting major sound events in an audio signal while shorter windows may be used in case additional sound events are needed to be detected. In this paper, the optimum length of the analysis window has been determined by trials and errors on the two (development) training data sets, D1 and D2. Figure 3 shows the phase slopes from a short excerpt of a music sample from D2 computed with three different analysis window lengths. The optimum analysis window was found to be 0.2s, thus slightly shorter than the shortest considered signal period (i.e., 0.3 s). A typical window like Hanning was used while the step size of the analysis was set to 5.8ms which corresponds to a sample rate, f_l, of 172 Hz for the onset signal, as suggested in [2].

Two further refinements of the approach as explained in Section 2 have been found to be necessary for music signals. The first is a zero crossing selection. In Figure 3 it can be observed that the shorter the analysis window the more positive zero-crossings are detected. It was observed that accepting only the zero crossings that are surrounded by large oscillations improves the accuracy on the development sets. Such oscillations can easily be detected by thresholding, as shown by the dotted lines in each sub plot of Figure 3. The positive threshold has been determined by the mean of the absolute values of the phase slope for a whole sample; the negative threshold is simply the negative of this value. A zero-crossing is selected if the minimum and the maximum amplitude of the phase slope function, before and after the zero-crossing, respectively, pass the corresponding thresholds. For example, the zero crossing at sample 800 in the middle plot was rejected, because it is not followed by a large maximum.

The second refinement is the usage of a multi band analysis. Dividing the spectrum into a number of bands has been shown to be meaningful for beat tracking [12, 13]. For this, the spectrum has been divided into four equally sized bands on logarithmic frequency scale. In each band, an onset detection using the phase slope method was performed. In order to get a single vector representation, the four band-wise onset signals, $\mathbf{y}_c[n], c = 1...4$, have been fused in the same

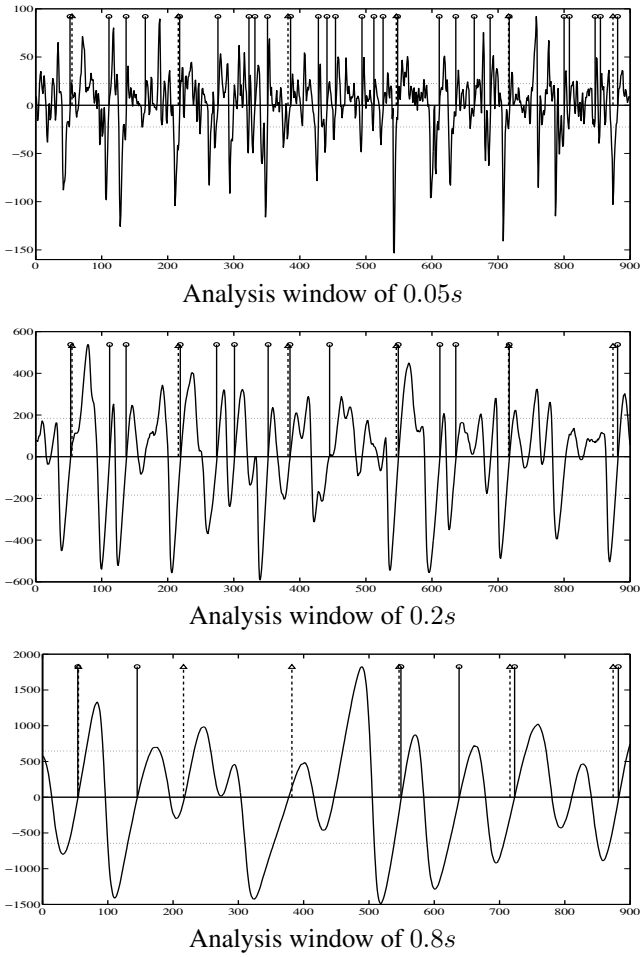

Figure 3. Influence of the analysis window length on the phase slope of a music sample from D2 (x-axis: samples, y-axis: phase slope amplitude, onsets: bold peaks with circle markers, annotation: dashed peaks with triangle markers, threshold for zero crossing selection: dotted lines)

way as in [2]:

$$\mathbf{y}[n] = \sum_{c=1}^{4}(6-c)\mathbf{y}_c[n] \tag{5}$$

giving more weight to lower bands. Indeed, the subband splitting as well as the fusion of bands have been found to improve the performance of the beat tracker as measured by (4).

3.2 Beat tracking

For the estimation of beat times from the band-wise onset signals an algorithm based on the method proposed by Klapuri et al. in [2] has been used. The algorithm had to be adapted to the type of onset signals that are obtained using the phase slope function. This modified beat tracker will be referred to as M-KLAP in the rest of the paper. Experiments for the development of M-KLAP have been conducted using the music data set, D2, described in Section 3.1.

3.2.1 Beat Period

For beat period estimation, Klapuri et al. suggest the computation of comb filter responses on each of the four bands separately, and summing afterwards. In M-KLAP, the band wise onset signals, $\mathbf{y}_c$, are simply summed using (5). Afterwards, the obtained onset vector is weighted with the sum of the spectral flux at each sample n:

$$\mathbf{y}_{flux}[n] = \mathbf{y}[n]\sum_{\omega} HWR(|X(\omega,n)| - |X(\omega,(n-1))|) \tag{6}$$

where HWR denotes a half wave rectification and $X(\omega, n)$ denotes the (short time) Fourier transform of the signal as used in the group delay computation in (2). Weighting the detected onsets in this way slightly but consistently improves performance.

The sample autocorrelation of the vector $\mathbf{y}_{flux}[n]$ is then computed in a rectangular window of $t_{win} = 8s$ length with a step size of one second. The maximum lag considered is $4s \times f_l$, which is equal to 688, since $f_l = 172Hz$. The centers of the analysis windows are positioned at times $[1s, 2s, ..., T_N]$, where $T_N = \lfloor N/f_l \rfloor$, zero padding has been applied. In the following, the beat periods β have been estimated using a *Hidden Markov Model*(HMM) as described in [2], where the beat period is referred to as tactus period. This results in a sequence of beat period estimations $\beta[k]$, with $k = 1...T_N/s$. The only change in the HMM is the use of flat priors for the beat periods, and that the beat periods do not depend on the simultaneous measure periods ((24) in [2]) in the Viterbi algorithm.

3.2.2 Beat Phase

In the phase estimation of the beat pulse ((27) in [2]), the computation of the likelihood of a phase $\Phi[k]$ in analysis frame k has been changed to

$$P(\hat{\mathbf{r}}_{\tilde{\mathbf{y}}\mathbf{y}}|\Phi[k]=l) = \sum_{c=1}^{4}(6-c)\sum_{n=0}^{8f_l}\tilde{\mathbf{y}}_k[n+l]\mathbf{y}_c[kf_L+n-4f_l] \tag{7}$$

where $\tilde{\mathbf{y}}_k$ is a reference pulse train of $t_{win}f_l + 1$ samples length, having an impulse at the middle position and a period equal to $\beta[k]$. Thus, just like in the estimation of the beat period, an eight second length window has been used. The weighted sum of the band wise correlations as computed in (7) is then used in an HMM framework as suggested in [2]. Again, incorporating spectral flux as in (6) has been tried, combined with an impulse selection instead of a Viterbi in order to get the final beat pulse. This would

have the advantage of avoiding the relatively long (8s) analysis window. However, results were slightly inferior to those achieved using the Viterbi algorithm. Because of this, this improvement was postponed for now. Note that the accuracy of measure and tatum periods [2] have not been evaluated, as the focus is the derivation of the beat information.

4 EXPERIMENTS

This Section compares the performance of the system as suggested by Klapuri et al.[2], denoted as KLAP, with the performance phase slope detected onsets as input to the modified beat tracker, which will be referred to as PS/M-KLAP. Two data sets of beat annotated pieces of music have been used for evaluation. The first has been used as a training set for the MIREX 2006 Audio Beat Tracking task [3], and consists of twenty 30 second excerpts from popular music songs. Each song has been beat annotated by several listeners, who were asked to tap the beat of the piece of music. In the following, this data set is referred to as T1. The second data set, (T2), consists of twenty 30 second excerpts from pieces of traditional Cretan music, downloaded from the Institute of Mediterranean Studies [4]. The beat for these pieces has been annotated by the first author. In contrast to T1, none of the songs contain percussive instruments, but only string instruments and vocals. It is worth to note, that none of the mentioned data sets have been used to find the optimal parameters of the system. For this purpose, only the development sets, D1 and D2, mentioned in Section 3, have been used.

As detailed in [2], the most appropriate estimator for the performance of a beat tracking system is the length of the longest continuous correct estimated section of the song, divided by the duration of the whole song. For example, for 30s duration of a song and 12s to be the longest continuously correct beat estimation duration, the accuracy is 40%. Furthermore, the beat estimation is judged as correct when its period is half or double the period of the annotation as well. A deviation of 0.175 times the annotated period length is tolerated. Note that a beat pulse train with the same period as the annotation pulse train is considered as incorrect whenever it has a half period offset (*off-beat*). Accuracies measured with this method will be referred to as A_{cont}. For convenience, also the accuracies as computed by (4) will be shown, denoted as A_{mir}, in order to be able to compare with scores achieved at the MIREX contest.

4.1 Proof of concept

In this Section, the KLAP and PS/M-KLAP beat trackers are applied to D1, the development set containing artificial signals. For each TNR level, the accuracies of the two beat tracking systems have been computed using (4) for all the signal periods in D1 (0.3s to 1s). Then the mean values and the standard errors have been computed for each TNR level. The mean accuracy values along with their corresponding standard errors, shown as error bars, are depicted in Figure 4. Without the addition of noise both approaches estimate a beat pulse train that is perfectly correlated with the position of the impulses in the signal. When the TNR decreases, the presented approach PS/M-KLAP is persistently more accurate. This proofs the hypothesis, that using the proposed approach, beat tracking will be more robust against noise which is important if an audio recording is noise corrupted. Also the presence of noise makes some of the possible percussion components found in music to soften. Based on the above results we expect the proposed approach to be also appropriate for musical signals without strong percussive components.

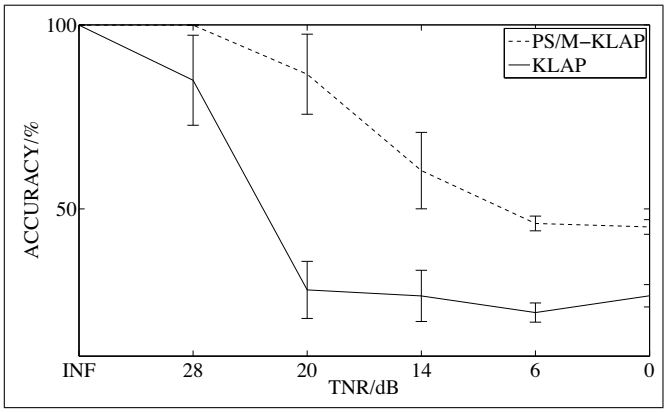

Figure 4. Accuracy of the beat tracking using the proposed method (PS/M-KLAP) and the algorithm of [2] (KLAP), on artificial signals of varying TNR

4.2 Results

The accuracies of the beat trackers applied to the music data sets T1 and T2 are depicted in Tables 1 and 2 for the accuracy measures A_{cont} and A_{mir}, respectively. On T1, the KLAP beat tracker is superior. The advantage of using the phase slope in the proposed way is distinct on T2. Here, the improvement compared to the state-of-the-art approach is 82%. This shows that using the proposed method, beat tracking in a signal with weak or no percussive content can be improved, while approaches using the amplitude information clearly fail on this data. Since the difference between the KLAP and PS/M-KLAP system, doesn't solely rely on the onset detection approach (there are modifications in beat tracking approach as well), it may be assumed that the differences in accuracy cannot be solely attributed to the onset detection method. To check this we decided to provide as input to the M-KLAP system the standard input of the KLAP

[3] http://www.music-ir.org/mirex/2006/index.php/Audio_Beat_Tracking
[4] http://gaia.ims.forth.gr/portal/

	PS/M-KLAP	KLAP
T1	54.3(0.058)	58.4(0.063)
T2	48.0(0.171)	26.3(0.122)

Table 1. Accuracies A_{cont} of the beat tracking on the two data sets, mean value/%(variance)

	PS/M-KLAP	KLAP
T1	45.0(0.028)	50.0(0.026)
T2	39.3(0.071)	21.4(0.085)

Table 2. Accuracies A_{mir} of the beat tracking on the two data sets, mean value/%(variance)

system, *i.e.* that derived from spectral flux [2]. For T1 and T2 data sets, the obtained results are $48\%/43.1\%$ and $22\%/27.8\%$, respectively for A_{cont}/A_{mir} measures. This shows the importance of the phase slope function in onset detection and in the context of beat-tracking. The slightly lower performance of PS/M-KLAP in T1 as compared to the standard system (KLAP) may be attributed to the implementation of beat tracking and we expect to further improve that part in the near future.

5 CONCLUSIONS

In this paper a new method to detect onsets using the average of the group delay was introduced and evaluated in a beat tracking framework. Advantages are the immediate detection of the onsets by locating the positive zero crossings of the phase slope, and the robustness for signals with little percussive content as shown in the experiments. The next steps to improve the method are a more efficient implementation of the phase slope computation and refinements of the beat tracker. Also, evaluation on different data sets of folk music will be performed.

6 ACKNOWLEDGEMENTS

The authors would like to thank Anssi Klapuri for providing his meter estimation code.

7 REFERENCES

[1] Daniel P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[2] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Acoustics Speech and Signal Processing*, in press.

[3] Simon Dixon, "Mirex 2006 audio beat tracking evaluation: Beatroot," in *MIREX at 7th International ISMIR 2006 Conference*, 2006.

[4] M. E. P. Davies and M. D. Plumbley, "Tempo estimation and beat tracking with adaptive input selection," in *MIREX at 7th International ISMIR 2006 Conference*, 2006.

[5] Simon Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects*, 2006.

[6] Dan Stowell and Mark Plumbley, "Adaprive whitening preprocessing applied to onset detectors," in *MIREX at 8th International ISMIR 2007 Conference*, 2007.

[7] Matthew E. P. Davies and Mark D. Plumbley, "Context-dependent beat tracking of musical audio," *Audio, Speech, and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, vol. 15, no. 3, pp. 1009–1020, March 2007.

[8] Andre Holzapfel and Yannis Stylianou, "Rhythmic similarity of music based on dynamic periodicity warping," in *ICASSP 2008*, 2008.

[9] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 325–333, 1995.

[10] V. Kandia and Y. Stylianou, "Detection of clicks based on group delay," *Accepted in Canadian Acoustics*, 2008.

[11] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, *Discrete-Time Signal Processing*, Prentice Hall, 1998.

[12] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 588–601, 1998.

[13] M. Goto and Y. Muraoka, "Music understanding at the beat level: Real-time beat tracking for audio signals," in *Proceedings of IJCAI 95 Workshop on Computational Auditory Scene Analysis*, 1995, pp. 68–75.