A SPECTRAL DIFFERENCE APPROACH TO DOWNBEAT EXTRACTION IN MUSICAL AUDIO

Matthew E. P. Davies and Mark D. Plumbley

Centre for Digital Music, Queen Mary, University of London
Mile End Road, London, E1 4NS, United Kingdom
matthew.davies@elec.qmul.ac.uk

ABSTRACT

We introduce a method for detecting downbeats in musical audio given a sequence of beat times. Using musical knowledge that lower frequency bands are perceptually more important, we find the spectral difference between band-limited beat synchronous analysis frames as a robust downbeat indicator. Initial results are encouraging for this type of system.

1. INTRODUCTION

Numerous approaches exist for the problem of beat tracking (e.g.[1, 2, 3, 4]), that of replicating the human ability of tapping in time to music. However much less attention has been given to higher level metrical analysis. One such problem is the extraction of downbeats from musical audio i.e. finding the *first* beat of each bar.

A robust downbeat extractor could be of considerable use within the context of music information retrieval: to enable fully automated rhythmic pattern analysis for genre classification [5]; to indicate likely temporal boundaries for structural audio segmentation [6]; and to improve the robustness of beat tracking systems by applying higher level knowledge [7].

The principal difficulty appears not in finding the number of beats per bar, the *time-signature*, but resolving the phase of the bar-level periodicity [7]. While this might appear a simple task, very few techniques have been found effective for solving this particular problem.

Goto [2] presents two approaches to downbeat estimation: for percussive music, automatically detected kick and snare drum events are compared to pre-defined rhythmic template patterns; for non-percussive music, short-term spectral frames (band-limited to 1kHz) are peak-picked and then histogrammed into beat length segments, where chord changes are used to infer higher level metrical structure. The two methods are combined within a single rhythm tracking system [2] which is shown to be highly accurate and operates in real-time. Goto's system however, has only been fully tested on a popular music database and restricted to music in 4/4 time with a constant tempo between 61 and 120 beats per minute (bpm).

Klapuri, Eronen and Astola [7] propose a meter tracking system which uses comb filter analysis within a probabilistic framework to simultaneously track three metrical levels: the *tatum*, *tactus* and measure. The phase of measure-level events, i.e. downbeats, are identified by matching rhythmic pattern templates to a mid-level representation calculated in four parallel sub-bands, where most emphasis is given to the lowest of these bands. Klapuri et al present results over a more varied test database than Goto's algorithm [2] and include cases which exhibit tempo variation. We therefore consider this approach the current state of the art for downbeat estimation.

In this paper we introduce a spectral difference approach to downbeat estimation. Although related to Goto's approach [2], we propose that percussive events and harmonic change can be used implicitly within a single spectral representation to infer downbeats. We require a sequence of beat times and the time-signature of the input signal to be known a priori – both of which are detected within our previously developed beat tracking system [1]. We partition an input signal into band-limited beat length frames and use the musical knowledge that lower pitched events are perceptually more important [4] by preserving spectral information within the range 0–1.4kHz. We calculate the Kullback-Leibler divergence between successive beat frames to form a spectral difference function. Downbeats are selected as those beats which globally lead to most spectral change.

We evaluate our downbeat model against that of Klapuri et al [7], with initial results indicating better performance for our model. However, current analysis is restricted to cases where the time-signature does not change and the tempo is approximately constant.

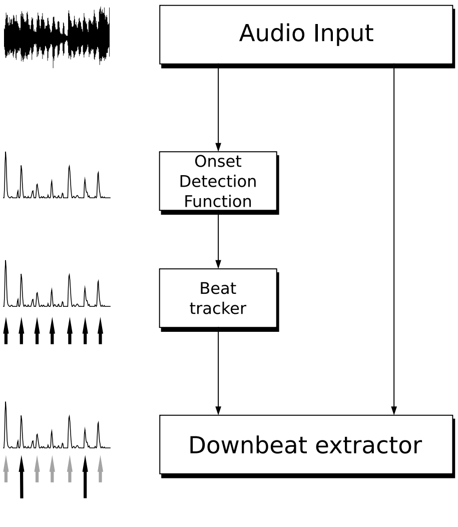

Figure 1: Overview of downbeat extraction model

The remainder of this paper is structured as follows. In section 2 we describe our approach to downbeat extraction. Section 3 contains results from an objective and subjective evaluation of our system with discussion and conclusions in sections 4 and 5.

2. APPROACH

Our approach requires that the beat locations and time signature of the input signal are known a priori. We use the output of our beat tracker as the front end to our system (fig 1), and provide a brief overview in the following section. A more detailed description can be found in [1].

2.1 Beat Tracker Overview

The first stage in the beat tracking algorithm is the transformation of an input audio signal (fig 2(a)) into a mid-level representation from which beat times can be robustly identified. We use the *complex spectral difference* onset detection function [8] (fig 2(b)). A sequence of beat times γ_m is recovered by passing the autocorrelation function of the detection function through a shift-invariant comb filterbank to extract the beat period. This is then used to identify the phase of the beats by cross-correlating the detection function with an impulse train with impulses at beat period intervals.

We classify the time-signature τ of the input by comparing the energy at integer multiples of the beat period in the autocorrelation function of the detection function

$$r(2\tau_b)+r(4\tau_b) > r(3\tau_b)+r(6\tau_b) \qquad (1)$$

where r is the autocorrelation function and τ_b is the beat period. If the above condition holds, we infer duple time and set $\tau = 4$, else we assume triple time and set $\tau = 3$.

Our model is able to follow local expressive timing variations and detect tempo changes, and uses context-dependent information to enforce contextual continuity with a single tempo hypothesis. Results in [1] indicate comparable beat tracking performance to the current state of the art [7].

2.2 Detecting Downbeats

Given the beats and time-signature, we partition an input signal $x(n)$ sampled at $f_s = 44.1$kHz into beat length segments $x_m(n)$. To retain the perceptually important lower end of the spectrum we resample the beat segments at $f_2 = (f_s/16) \approx 2.8$kHz. Experiments suggest that the precise spectral range is not critical, so for convenience we downsample the audio by a factor of 16. We then find the spectrum $X_m(\omega)$ of the m^{th} band-limited beat segment

$$X_m(\omega) = \frac{1}{N}\sum_{n=1}^{N/2} w(n)x_m(n)e^{-j\frac{2\pi n\omega}{N}} \qquad (2)$$

where, to account for varying beat length segments, we fix $N = 512$. To reduce the contribution of the least significant spectral components we apply an adaptive threshold. The threshold is defined as the convolution of the magnitude spectrum of the m^{th} beat frame $|X_m(\omega)|$ with an empirically derived smoothing kernel,

$$H(\omega) = \begin{cases} 0.2 & \omega = 1,\ldots,5 \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

This is then subtracted from $|X_m(\omega)|$ to leave a modified spectral frame $\hat{X}_m(\omega)$ (fig 2(c))

$$\hat{X}_m(\omega) = |X_m(\omega)| - |X_m(\omega)| * H(\omega) \qquad (4)$$

$*$ refers to the convolution operator. $\hat{X}_m(\omega)$ is half-wave rectified to set any negative valued elements to zero.

We derive a spectral difference function $D(m)$ using the Information Theoretic measure, Kullback-Leibler (K-L) divergence [9] (fig 2(d)). To ensure a real-valued output we add a negligible non-zero constant to each spectral frame and then normalise it to sum to unity

$$D(m) = \sum_{\omega=1}^{N/2} \hat{X}_m(\omega)\log_e \frac{\hat{X}_m(\omega)}{\hat{X}_{m+1}(\omega)} \qquad (5)$$

Initial experiments showed K-L divergence to be more effective than using Euclidean distance, a result also observed by Hainsworth and Macleod [10] who detect harmonic change for note onset detection.

Given the number of beats per bar τ, we calculate a signal $\eta(\varphi)$ as the measure of spectral change at each downbeat candidate $\varphi = 1,\ldots,\tau$

$$\eta(\varphi) = \sum_{m=1}^{M} D(\tau(m-1)+\varphi) \qquad (6)$$

where M is the number of complete bar length segments. We then extract the beat leading to most spectral change φ_d as the index of the maximum value in $\eta(\varphi)$

$$\varphi_d = \arg\max_{\varphi} \eta(\varphi) \qquad (7)$$

Assuming a steady tempo, we may then extract the downbeats γ_d from the beat indices γ_m, by setting $d = (m-1)\tau + \varphi_d$.

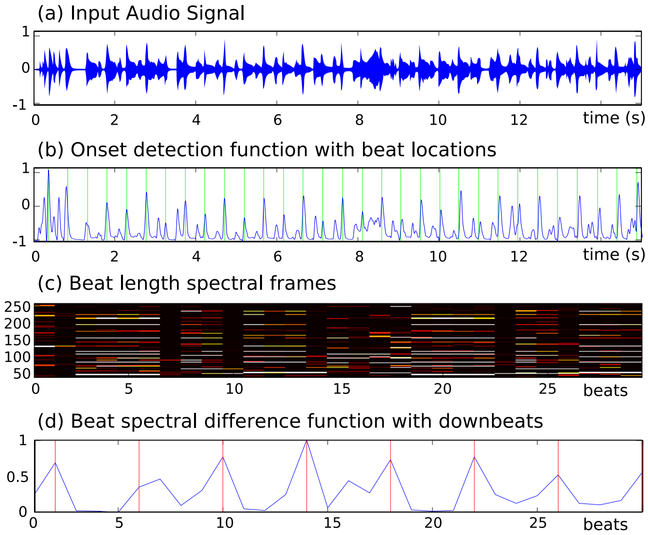

Figure 2: Top to bottom: (a) input audio signal; (b) onset detection function with vertical lines indicating beat locations ; (c) band-limited spectrum beat frames and (d) spectral difference function with extracted downbeat indices.

3. RESULTS

We evaluate our approach to downbeat estimation in two stages. First we present results from objective analysis, where extracted downbeats are compared to manually annotated values from a test database [3]. We then perform a subjective evaluation (on cases where the beat tracking is

Downbeat Algorithm	Accuracy (181 files)	Accuracy (72 files)
KEA	40.8 %	69.9 %
DP B	52.6 %	81.2 %
DP A	72.4 %	76.4 %

Table 1: Results for proposed approach using manual annotations: (DP A), beat tracker output: (DP B) and the Klapuri et al [7] algorithm: (KEA). Column 2 contains results across the full 181 file database, with results for a subset of 72 accurately beat tracked files in column 3.

accurate) to characterise the types of errors made by the system. In both stages we compare the output of our algorithm in a fully automatic setting, using our beat tracker [1] to provide the beat indices, labelled DP B, with a semi-automatic approach DP A which uses the manual beat annotations. We also include results from the Klapuri et al model [7], which we refer to as KEA.

3.1 Objective Analysis

The objective approach to evaluation compares the output downbeat indices γ_d to downbeats a_j annotated by a trained musician from a beat tracking test database [3]. The complete database contains 222 files across six musical genres (rock, dance, jazz, folk, classical and choral). Each excerpt is between 30 and 60 seconds, mono and sampled at 44.1kHz with 16 bit resolution. We listened to the annotations for each example and removed those cases where the downbeat annotations were ambiguous (predominantly the classical and choral examples) or contained changes in time-signature or tempo. We retained a total of 181 files upon which we tested our algorithm.

For the automatic approaches (DP B and KEA) we define a downbeat γ_d to be accurate if it falls within a specified allowance window around the annotated value a_j, such that

$$a_j - \theta\Delta_j^- < \gamma_d < a_j + \theta\Delta_j^+ \qquad (8)$$

where Δ_j^- and Δ_j^+ are the previous and subsequent inter-annotation intervals and θ is the allowance window. In these experiments we set $\theta = 0.1$ as used by Klapuri et al in their recent study [7]. We recognise that when using the annotated beats and time-signature for DP A, the accuracy for each file will either be 0 or 100%, results are included to indicate an upper limit for accuracy of DP B. Results comparing DP B, DP A and KEA are shown in column 2 of Table 1.

An accuracy of 72.4% for the proposed system DP A suggests that spectral difference is an appropriate measure for extracting the downbeat. The overall accuracy is lower for the fully automatic system DP B. This is as expected, since errors in beat tracking and time-signature extraction are naturally carried over. However our approach is still more successful than that of Klapuri et al [7].

An intuitive observation from the results across the 181 file database is that the downbeat accuracy is only as good as the beat tracking performance. To further analyse the automatic approaches we extracted a subset of files from the test database retaining only those in which our algorithm and Klapuri et al's [7] were found to be *both* 95% accurate in beat tracking (further details comparing the beat tracking performance can be found in [1]). Results shown in column 3 of Table 1, confirm that accurate beat tracking significantly increases downbeat accuracy.

3.2 Subjective Analysis

An unexpected outcome of the objective evaluation (shown in Table 1) revealed the downbeats to be more accurate in the fully automatic case (DP B) than when using the annotated data (DP A), 81.2% compared to 76.4%. This was confirmed through subjective audition of the 72 subset files in which we identified accurate downbeat assignment in 55, 61 and 53 instances for the DP A, DP B and KEA respectively. A graphical representation of the successes and failures of the models are shown in fig 3, where '1' indicates a correct downbeat, '2','3' and '4' represent the perceived location of erroneous downbeats and 'W' refers to an incorrect time-signature estimate.

Across many cases we also observed perceptually more consistent beat timing for the automatic models than for the annotated data, most evident in cases where the tempo was constant. From this we infer that inaccuracies in the localisation of the beats (the result of human annotation) are significant when comparing beat length analysis frames, as spectral changes can become blurred across beat boundaries, making them harder to detect, thus explaining why DP A was less successful than DP B. We intend to investigate this in greater detail within our future work.

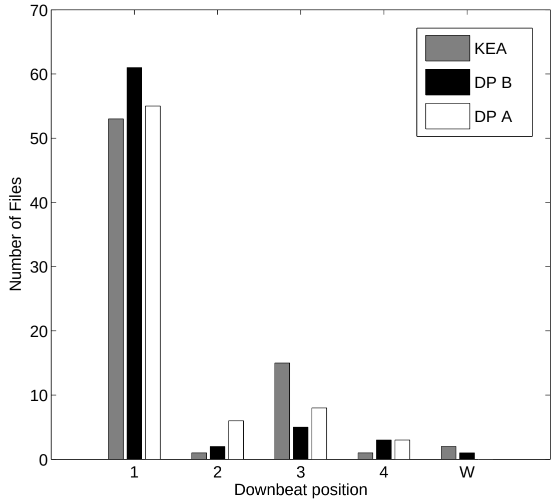

Figure 3: Subjective classification of downbeat locations. KEA - Klapuri et al [7], DP B - proposed model using beats, DP A - proposed model using annotations. W refers to incorrect time signature.

4. DISCUSSION

Our initial results have shown that, for the test database used, our spectral difference approach is able to correctly infer the downbeat more reliably than Klapuri et al's [7] rhythmic template approach. We should note, however, that the most common error made by the KEA method is in selecting the downbeat as the '3' rather than the '1', i.e. the downbeat offset by two beats. This behaviour is not replicated by our approach which has a more uniform distribution of bad downbeat choices (fig 3). While selecting the '3' is clearly not correct, it is potentially more desirable than picking either '2' or '4', as this is more likely to correspond to a strong beat than a weak beat. This suggests that a combination of our approach with the KEA model [7] might yield better performance than either one individually, using the KEA model to reliably select either the '1' or '3' beats followed by our spectral difference approach to discern which is the most likely downbeat. At present this is left as a topic for future work.

In addition to a possible combined rhythmic pattern approach, we plan to extend our model to overcome its present limitations – most notably that any omitted beats or changes in tempo or time-signature cause errors from which our proposed downbeat extraction model cannot recover. To allow for variation in tempo and time-signature, we intend to track spectral changes within the spectral difference function, replacing the current 'winner takes all' strategy where a single downbeat candidate is used to extract all bar lines. We also plan to investigate methods for deriving the time-signature automatically from the spectral difference function. In a similar approach to Gouyon and Herrera [11] we could take the autocorrelation function of the spectral difference signal and detect the time-signature as the beat lag with highest energy. This would then enable the analysis of songs in time-signatures other than 3/4 or 4/4. Deriving the time-signature directly from the beat spectral difference function in this way, would also allow our downbeat estimator to operate directly with the output of any beat tracking system.

In evaluating our downbeat model, we plan to incorporate a larger test database in addition to comparing performance against other published approaches to downbeat extraction. These include the approaches of Jehan [12] and Allan [13] who present notably different formulations of the problem to those of Goto [2] and Klapuri et al [7]. Jehan [12] provides style specific training data to his model which is able to predict downbeat locations without the need for a beat tracker. Allan [13] has a more generic approach to metrical analysis. Given manual annotations at one metrical level, he uses spectral similarity of varying segment lengths to extract spectral patterns and infer the metrical level above that of the annotations. His approach can be used recursively from the sub-beat level to give bar lines and segment boundaries.

5. CONCLUSIONS

We have presented a simple algorithm for the extraction of downbeats in musical audio. We evaluate our algorithm by comparing extracted bar lines to human annotated data with an overall accuracy of 53% rising to 81% for cases where beat tracking is accurate. Under both conditions our approach is more accurate than a reference state of the art meter analysis system.

6. ACKNOWLEDGEMENTS

Special thanks to Anssi Klapuri for making his source code available, to Hamish Allan for invaluable technical discussions and to the anonymous reviewers for their helpful suggestions.

Matthew Davies is supported by a college studentship. This research has been partially partially supported by EPSRC grants GR/S75802/01, GR/S82213/01 and by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents). http://www.semanticaudio.org.

REFERENCES

[1] M. E. P. Davies and M. D. Plumbley, "Beat tracking with a two state model," in *Proceedings of ICASSP*, Philadelphia, PA, USA. March 18–23, 2005

[2] 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, 2001

[3] S. W. Hainsworth, *Techniques for the automated analysis of musical audio*, Ph.D Thesis, Department of Engineering, Cambridge University, 2004

[4] S. Dixon and E. Cambouropoulos, "Beat Tracking with Musical Knowledge," in *Proceedings of the 14th European Conference on Artificial Intelligence*, ed. W.Horn, IOS Press, Amsterdam. pp. 626–630. 2000

[5] S. Dixon, F. Gouyon and G. Widmer, "Towards characterisation of music via rhythmic patterns," in *Proceedings of ISMIR 2004*, Barcelona, Spain. pp. 509–516. 2004

[6] M. Levy, M. Sandler and M. Casey, "Extraction of high level musical structure from audio data and its application to thumbnail generation," *to appear ICASSP 2006*

[7] A. Klapuri, A. Eronen and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Speech and Audio Processing* vol 14, no. 1, 2006 *to appear*

[8] 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

[9] S. Kullback and R. A. Leibler, "On information sufficiency," *The Annals of Mathematical Statistics* vol. 22, no. 1, pp. 79–86, March 1951

[10] S. Hainsworth and M. Macleod, "Onset detection in musical audio signals," in *Proceedings of ICMC*, Singapore, September 2003

[11] F. Gouyon and P. Herrera, "Determination of the Meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *Proceedings of Audio Engineering Society, 114th Convention*, Amsterdam, The Netherlands. 2003

[12] T. Jehan, "Downbeat prediction by listening and learning," in *Proceedings of WASPAA*, Mohonk, NY, USA. pp. 267–270, 2005

[13] H. Allan, *Bar lines and beyond - Metre tracking in digital audio*, MSc Thesis, School of Informatics, University of Edinburgh. 2004