BETTER BEAT TRACKING THROUGH ROBUST ONSET AGGREGATION

Brian McFee

Center for Jazz Studies
Columbia University
brm2132@columbia.edu

Daniel P.W. Ellis

LabROSA, Department of Electrical Engineering
Columbia University
dpwe@columbia.edu

ABSTRACT

Onset detection forms the critical first stage of most beat tracking algorithms. While common spectral-difference onset detectors can work well in genres with clear rhythmic structure, they can be sensitive to loud, asynchronous events (*e.g.*, off-beat notes in a jazz solo), which limits their general efficacy. In this paper, we investigate methods to improve the robustness of onset detection for beat tracking. Experimental results indicate that simple modifications to onset detection can produce large improvements in beat tracking accuracy.

***Index Terms*—** Music information retrieval, beat tracking

1. INTRODUCTION

Beat-tracking — the detection of "pulse" or salient, rhythmic events in a musical performance — is a fundamental problem in music content analysis. Automatic beat-detection methods are often used for chord recognition, cover song detection, structural segmentation, transcription, and numerous other applications. A large body of literature has developed over the past two decades, and each year sees numerous submissions to the Music Information Retrieval Evaluation eXchange (MIREX) beat tracking evaluation [1].

A common general strategy for beat tracking operates in two stages. First, the audio signal is processed by an onset strength function, which measures the likelihood that a musically salient change (*e.g.*, note onset) has occurred at each time point. The tracking algorithm then selects the beat times from among the peaks of the onset strength profile.

As we will demonstrate, the behavior of standard onset detectors tends to be dominated by the loudest events, typically produced by predominant or foreground instruments and performers. In many styles of western, popular music — *e.g.*, rock, dance, or pop — this presents no difficulty. Often, the beat is unambiguously driven by percussion or foreground instrumentation, resulting in clear rhythmic patterns which are amenable to signal analysis.

The assumption that beat derives from the predominant foreground instrumentation does not hold in general across diverse categories of music. As a concrete example, a soloist in a jazz combo may play a syncopated rhythm, or off-beat for aesthetic or expressive purposes, while the accompaniment maintains a steady pulse in the background. In such cases, we would hope that a beat tracker would adaptively tune out the foreground instrumentation and focus on the rhythmically salient portion of the signal.

Reliable detection and separation of rhythmic elements in a recording can be quite difficult to achieve in practice. Humans can tap along to a performance and adapt to sudden changes in instrumentation (*e.g.*, a drum solo), but this behavior is difficult for an algorithm to emulate.

1.1. Our contributions

In this work, we investigate two complementary techniques to improve the robustness of beat tracking and onset detection. First, we propose across-frequency median onset aggregation, which captures temporally synchronous onsets, and is robust to spurious, large spectral deviations. Second, we examine two spectrogram decomposition methods to separate the signal into distinct components, allowing the onset detector to suppress noisy or arrhythmic events.

1.2. Related work

Onset detection is a well-studied problem in music information retrieval, and a full summary of recent work on the subject lies well beyond the scope of this paper. Within the context of beat-tracking, the surveys by Bello *et al.* [2] and Collins [3] provide general introductions to the topic, and evaluate a wide variety of different approaches to detecting onset events.

Escalona-Espinosa applied harmonic-percussive separation to beat-tracking, and derived beat times from the self-similarity over features extracted from the different components [4]. The approach taken in this work is rather different, as we evaluate onset detectors derived from a single component of a spectrogram decomposition.

Peeters [5] and Wu *et al.* [6] highlight tempo variation as a key challenge in beat tracking. While tempo variation is

This work was supported by a grant from the Mellon foundation, and grant IIS-1117015 from the National Science Foundation (NSF).

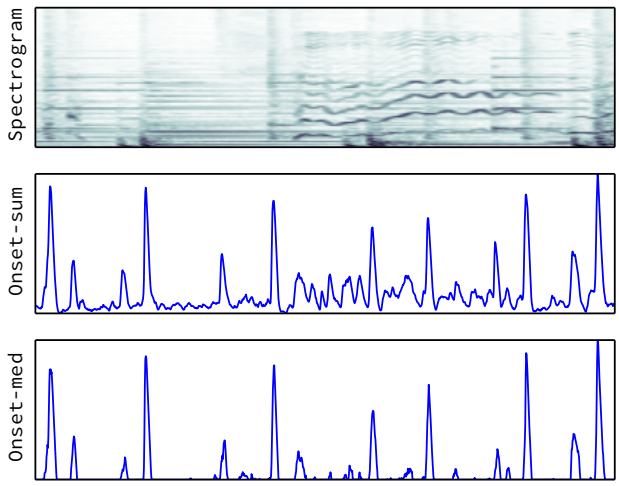

Fig. 1. An example spectrogram (top) derived from five seconds of vocals, piano, and drums. Sum across frequency bands to derive onset strength (middle) results in spurious peaks due to pitch bends and vibrato. Median aggregation (bottom) produces a sparser onset strength function, and retains the salient peaks.

indeed a challenge, our focus here is on improving the detection of salient onset events; the tracking algorithm used in this work maintains a fixed tempo estimate for the duration of the track, but allows for deviation from the tempo.

Alonso *et al.* [7] and Bello *et al.* [2] propose using temporal median-filtering of the onset strength envelope to reduce noise and suppress spurious onset events. Temporal smoothing differs from the median-aggregation method proposed in this work, which instead filters across frequencies at each time step prior to constructing the onset envelope.

This article addresses the early stages of beat tracking. Rather than develop a new framework from scratch, we chose to modify the method proposed by Ellis [8], which operates in three stages:

1. compute an onset strength envelope $\omega(t)$,
2. estimate the tempo by picking peaks in the windowed auto-correlation of $\omega(t)$, and
3. select beats consistent with the estimated tempo from the peaks of $\omega(t)$ by dynamic programming.

Keeping steps 2–3 fixed allows us to evaluate the contribution to accuracy due to the choice of onset strength function. We expect that improvements to onset detection can be applied to benefit other beat tracking architectures.

2. MEDIAN ONSET AGGREGATION

The general class of onset detector functions we consider is based on *spectral difference*, *i.e.*, measuring the change in spectral energy across frequency bands in successive spectrogram frames [2]. The tracker of Ellis [8] uses the sum across bands of thresholded log-magnitude difference to determine the onset strength at time t:

$$\omega_s(t) := \sum_f \max(0, \log S_{f,t} - \log S_{f,t-1}), \qquad (1)$$

where $S \in \mathbb{R}_+^{d \times T}$ denotes the (Mel-scaled) magnitude spectrogram. This function effectively measures increasing spectral energy over time across *any* frequency band f, and its magnitude scales in proportion to the difference.

Note that ω_s can respond equally to either a large fluctuation confined to a single frequency band, or many small fluctuations spread across multiple frequency bands. The latter case typically arises from either a percussive event or multiple synchronized note onset events, both of which can be strong indicators of a beat. However, the former case can only arise when a single source plays out of sync with the other sources, such as a vocalist coming in late for dramatic effect.

To better capture temporally synchronous onset events, we propose to replace the sum across frequency bands with the median operator:

$$\omega_m(t) := \operatorname*{median}_f \max(0, \log S_{f,t} - \log S_{f,t-1}). \qquad (2)$$

This simple modification improves the robustness of the onset strength function to loud, asynchronous events.

As illustrated by Figure 1, the resulting onset envelope tends to be sparser, since it can only produce non-zero values if more than half of the frequency bins increase in energy simultaneously. [1] Consequently, pitch bends have a negligible effect on ω_m, since their influence is typically confined to a small subset of frequencies.

3. SPECTROGRAM DECOMPOSITION

In a typical musical recording, multiple instruments will play simultaneously. When all instruments (generally, sound sources) are synchronized, computing onsets directly from the spectrogram is likely to work well. However, if one or more sources play out of sync from each-other, it becomes difficult to differentiate the rhythmically meaningful onsets from the off-beat events. This motivates the use of source separation techniques to help isolate the sources of beat events. In this work, we applied two different source-separation techniques which have been demonstrated to work well for musical signals: harmonic-percussive source separation [9], and robust principal components analysis [10].

3.1. Harmonic-percussive source separation

Harmonic-percussive source separation (HPSS) describes the general class of algorithms which decompose the magnitude

[1] In preliminary experiments, alternative quantile estimators (25th and 75th percentile) were found to be inferior to median aggregation.

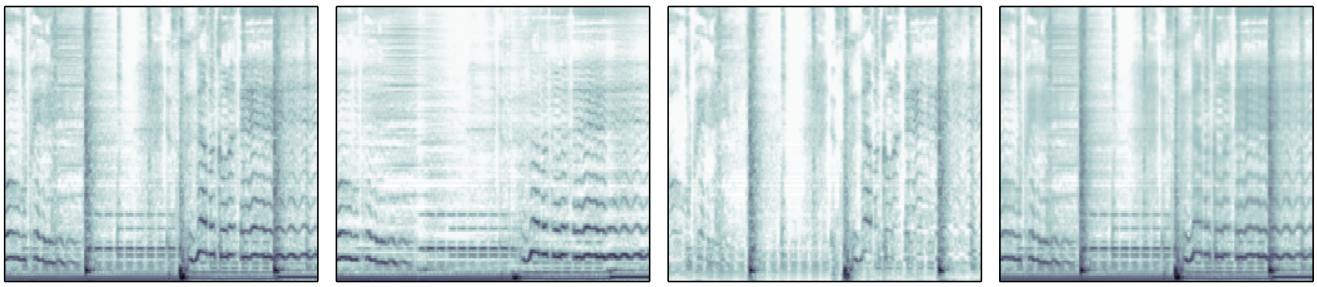

Fig. 2. Examples of spectrogram decomposition methods: (a) five seconds of a (Mel-scaled) spectrogram, consisting of guitar, bass, drums, and vocals; (b) the harmonic component emphasizes sustained tones (horizontal lines); (c) the percussive emphasizes transients (vertical lines); (d) the low-rank component retains harmonics and percussives, but suppresses vocal glides.

spectrogram as $S = H + P$, where H denotes *harmonics* — sustained tones concentrated in a small set of frequency bands — and P denotes *percussives* — transients with broad-band energy [9].

In this work, we used the median-filtering method of Fitzgerald [11]. Let η and π denote the harmonic- and percussive-enhanced spectrograms:

$$\pi := \mathcal{M}(S, w_p, 1)$$
$$\eta := \mathcal{M}(S, 1, w_h),$$

where $\mathcal{M}(\cdot, w_p, w_h)$ denotes a two-dimensional median filter with window size $w_p \times w_h$. The percussive component P is then recovered by soft-masking S:

$$P_{f,t} = S_{f,t} \left(\frac{\pi_{f,t}^p}{\pi_{f,t}^p + \eta_{f,t}^p} \right),$$

where $p > 0$ is a scaling parameter (typically $p = 1$ or 2). Given P, the harmonic component H is recovered by $H = S - P$.

Figure 2 (a–c) illustrates an example of HPSS on a short song excerpt. The harmonic component (b) retains most of the tonal content of the original signal (a), while the percussive component (c) retains transients.

In the context of beat tracking, it may be reasonable to use either H or P as the input spectrogram, depending on the particular instrumentation. While percussive instruments reliably indicate the beat in many genres (rock, dance, pop, *etc.*), this phenomenon is far from universal, particularly when the signal lacks percussion (*e.g.*, a solo piano).

3.2. Robust principal components analysis

In contrast to a fixed decomposition (*i.e.*, HPSS), it may be more effective to apply an adaptive decomposition which exploits the structure of the spectrogram in question. Recently, Yang demonstrated that robust principal components analysis (RPCA) can be effective for separating vocals from accompanying instrumentation [10, 12]. In this setting, RPCA finds a low-rank matrix $L \approx S$ which approximates S by solving the following convex optimization problem

$$L \leftarrow \operatorname*{argmin}_L \|L\|_* + \lambda \|S - L\|_1, \tag{3}$$

where $\|\cdot\|_*$ denotes the nuclear norm, $\|\cdot\|_1$ is the element-wise 1-norm, and $\lambda > 0$ is a trade-off parameter.

In practice, the low-rank approximation tends to suppress pitch bends and vibrato, which are both common characteristics of vocals and may account for some of its success at vocal separation. As shown in Figure 1, pitch bends can trigger spurious onset detections due to lack of temporal continuity within each frequency band, and should therefore be suppressed for beat tracking.

4. EVALUATION

To evaluate the proposed methods, we measured the alignment of detected beat events to beat taps generated by human annotators. Following previous work, we report the following standard beat tracking metrics [13]:

AMLt (range: $[0, 1]$, larger is better) is a continuity-based metric that resolves predicted beats at different allowed metrical levels (AML), and is therefore robust against doubling or halving of detected tempo;

F-measure (range: $[0, 1]$, larger is better) measures the precision and recall of ground truth beat events by the predictor;

Information gain (range: $[0, \cdot)$, larger is better) measures the mutual information (in bits) between the predicted beat sequence and the ground truth annotations.

Because different human annotators may produce beat sequences at different levels of granularity for the same track, meter-invariant measures such as AMLt and Information Gain are generally preferred; we include F-measure for completeness.

Algorithms were evaluated on SMC Dataset2 [14], which contains 217 40-second clips from a wide range of genres

and instrumentations (classical, chanson, blues, jazz, solo guitar, *etc.*). This dataset was designed to consist primarily of difficult examples, and represents the most challenging publicly available dataset for beat tracking evaluation. We include comparisons to the best-performing methods reported by Holzapfel *et al.* [14] — Degara *et al.* [15], Böck and Schedl [16], and Klapuri *et al.* [17] — and to the original implementation described by Ellis [8].

4.1. Implementation

Each track was sampled at 22050Hz, and Mel-scaled magnitude spectrograms were computed with a Hann-windowed short-time Fourier transform with 2048 samples ($\approx$ 93ms), hop of 64 samples ($\approx$ 3ms), $d = 128$ Mel bands, and a maximum frequency cutoff of 8000Hz. HPSS was performed with a hop of 512 samples, window sizes $w_p = w_h = 31$, and the power parameter was set to $p = 2.0$. Following Candès *et al.* [10], the RPCA parameter was set to $\lambda = \sqrt{T}$, where T denotes the number of frames. All algorithms were implemented in Python using `librosa`.[2]

4.2. Results

Table 1 lists the average scores achieved by the proposed methods on SMC Dataset2. For each metric, methods which achieve statistical equivalence to the best performance are listed in bold. Statistical significance was determined with a Bonferroni-corrected Wilcoxon signed-rank test at level $\alpha = 0.05$.

We first observe the gap in performance between *sum-Full* and Ellis [8], which differ only in their choice of parameters: the original implementation used a lower sampling rate (8000Hz), smaller window (256 samples) and hop (32 samples, 4ms), and fewer Mel bands ($d = 32$).[3] Except for the harmonic component method, all sum-based methods (first group of results) perform comparably well.

Replacing sum onset aggregation with median aggregation (second group of results) boosts performance uniformly: for each decomposition (except harmonic) and each metric, median aggregation only improves the score. The largest improvement is observed on the percussive component. Across all metrics, applying median aggregation to the percussive component ties for the highest score among all methods.

The RPCA method (Low-rank) did not yield significant improvements over either the full spectrogram or HPSS methods. This may be due to the fact that the dataset consists primarily of instrumental (even single-instrument) recordings, where there is less obvious benefit to source separation methods.

[2] `http://github.com/bmcfee/librosa/`

[3] The present implementation also includes a small constant timing correction, which improves performance for some metrics, but is known to not affect the information gain score [13].

Table 1. Beat tracker performance on SMC Dataset2.

Algorithm	AMLt	F-measure	Inf. gain
sum-Full	**0.290**	0.347	0.835
sum-Harmonic	0.222	0.283	0.655
sum-Percussive	**0.322**	**0.366**	0.858
sum-Low-rank	**0.300**	0.349	0.838
med-Full	0.340	**0.375**	**0.965**
med-Harmonic	0.224	0.268	0.720
med-Percussive	**0.366**	**0.383**	**1.005**
med-Low-rank	**0.347**	**0.376**	**0.965**
Böck & Schedl [16]	0.261	**0.401**	**0.928**
Degara *et al.* [15]	**0.334**	0.348	0.914
Ellis [8]	0.208	0.352	0.625
Klapuri *et al.* [17]	**0.339**	0.363	0.940

5. CONCLUSION

We evaluated two complementary techniques for improving beat tracking: onset aggregation, and spectrogram decomposition. The proposed median-based onset aggregation yields substantial improvements in beat tracker accuracy over the previous, sum-based method. Combining median onset aggregation with percussive separation results in the best performance on the SMC2 dataset.

6. ACKNOWLEDGMENTS

The authors acknowledge support from The Andrew W. Mellon Foundation, and NSF grant IIS-1117015.

7. REFERENCES

[1] J.S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[2] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler, "A tutorial on onset detection in music signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 1035–1047, 2005.

[3] Nick Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *Audio Engineering Society Convention 118*, 2005.

[4] Bernardo Escalona-Espinosa, "Downbeat and meter estimation in audio signals," *Master's Thesis, Technische Universität Hamburg-Harburg*, 2008.

[5] Geoffroy Peeters, "Time variable tempo detection and beat marking," in *Proc. ICMC*, 2005.

[6] Fu-Hai Frank Wu, Tsung-Chi Lee, Jyh-Shing Roger Jang, Kaichun K Chang, Chun Hung Lu, and Wen Nan Wang, "A two-fold dynamic programming approach to beat tracking for audio music with time-varying tempo," in *Proc. ISMIR*, 2011.

[7] Miguel Alonso, Bertrand David, and Gaël Richard, "Tempo and beat estimation of musical signals," in *Proc. International Conference on Music Information Retrieval*, 2004, pp. 158–163.

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

[9] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. ISMIR*, 2008, pp. 139–144.

[10] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11, 2011.

[11] Derry Fitzgerald, "Harmonic/percussive separation using median filtering," 2010.

[12] Yi-Hsuan Yang, "On sparse and low-rank matrix decomposition for singing voice separation," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 757–760.

[13] Matthew E.P. Davies, Norberto Degara, and Mark D Plumbley, "Evaluation methods for musical audio beat tracking algorithms," 2009.

[14] A. Holzapfel, M. E.P. Davies, J.R. Zapata, J.L. Oliveira, and F. Gouyon, "Selective sampling for beat tracking evaluation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2539–2548, 2012.

[15] Norberto Degara, Enrique Argones Rúa, Antonio Pena, Soledad Torres-Guijarro, Matthew EP Davies, and Mark D Plumbley, "Reliability-informed beat tracking of musical signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 290–301, 2012.

[16] Sebastian Böck and Markus Schedl, "Enhanced beat tracking with context-aware neural networks," in *Proc. Int. Conf. Digital Audio Effects*, 2011.

[17] Anssi P Klapuri, Antti J Eronen, and Jaakko T Astola, "Analysis of the meter of acoustic musical signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 342–355, 2006.