# IMPROVED DETECTION OF SEMI-PERCUSSIVE ONSETS IN AUDIO USING TEMPORAL REASSIGNMENT.

*Ken O'Hanlon, Mark B. Sandler*

Centre for Digital Music
Queen Mary University of London

## ABSTRACT

Onset detection is a fundamental task in musical signal processing, providing information for higher level applications. Different classes of onsets can be found in musical signals, determined as being hard, or soft, by the initial energy transfer. Most onset detectors are general purpose and attempt to detect both classes of onsets, although some specifically attempt to detect soft onsets. Temporal reassignment operators related to group delay have previously been employed in onset detectors for the purposes of soft onset detection and pruning of time-frequency elements deemed to consist of vibrato. We consider the use of temporal reassignment for the detection of hard onsets and also employ the second mixed derivative of phase as a means to prune the spectral energy. Experimental validation of the proposed approach is given, showing improvements relative to state-of-the-art general purpose onset detectors for the specific tasks.

***Index Terms***— Onset detection, music, reassignment

## 1. INTRODUCTION

Onset detection [1] is an important task in musical signal processing, enabling higher-level applications such as beat tracking [2], metric detection and modulation [3] and enhancing piano transcription [4]. Onset detection is generally divided into the separate tasks of hard onset and soft onset detection. Hard onsets consider percussive and semi-percussive instruments, which can be characterised by an increase in energy due to e.g. a drum or piano string being struck. On the other hand, soft onsets, such as might be more typical of woodwind or bowed string instruments may lack a sharp energy increase, and some specialised algorithms have been developed for their detection. Many algorithms try to incorporate elements that allow soft onset detection to be performed simultaneously to hard onset detection [1] [5]. Nonetheless there may be applications in which the detection of soft onsets is unnecessary. Audio stem formats, consisting of individual instrument tracks which allow enhanced user interactions and remixing, are becoming more popular. Furthermore, it has been shown that performing music analysis tasks on such multi-track data may improve upon processing the mixed data [6]. We consider that a hard onset detector may be desirable.

Different approaches have been taken to the design of onset detectors, however most fundamentally consider the differences between successive frames in spectrogram, or similar time-frequency representation [1]. The methods often differ by the features that are used, or measures between subsequent features [2], from which an onset detection function (ODF) is derived which is usually post-processed to estimate the onsets. Methods based on neural networks have been proposed [7], with promising results, however the same authors also propose a more traditional onset detection method known as superflux [5] that is generally considered state-of-the-art for a general purpose onset detection system.

Reassigned Spectrogram [8] are one of a class of methods, including the derivative [9] [10] and difference [11] methods that employ phase derivatives in a spectrogram to produce higher resolution in both time and frequency estimates, thereby reassigning the energy in a spectrogram to off-grid points. Such methods have previously been employed for chord recognition [12] and separation of percussive and low pitched signal elements [13]. Temporal reassignment has previously been employed for the purpose of onset detection. A variant of the Superflux method is the complex flux algorithm [14] which introduces a local group delay element that is used to further ameliorate the effects of vibrato. In [15] the group delay is employed for the detection of soft onsets, such as in flute music. In order to detect soft onsets multi-band estimation is performed. This work is expanded in [16] where the phase slope onset detector is fused with other onset detectors.

In this paper we propose a hard onset detector. Similar to [15] the group delay is employed to detect onset candidates, however in the context of hard onsets the multi-band approach employed in [15] is unnecessary. A further step is proposed that employs the mixed second derivative of the phase in order to mask the spectrogram, before using the masked spectral energy to weight the onset candidates. In the rest of this paper we first describe the reassignment method and onset detection in more detail before describing the proposed onset detector. Experimental results are then given which validate the proposed approach.

This research is funded by ESPRC Program Grant EP/L019981/1

## 2. BACKGROUND

### 2.1. Reassignment Method

The spectrogram is a time-frequency representation with coefficients assigned to a grid with equally spaced frequency bins and time frames. High resolution spectrograms localise each time-frequency element to a potentially off-grid point which may more accurately represent the energy found within the support of the time-frequency atom. Such reassignment can be performed both in the frequency

$$\hat{\omega}(\omega,\tau) = \omega + \frac{d\phi(\omega,\tau)}{d\tau} \tag{1}$$

and temporal domains

$$\hat{\tau}(\omega,\tau) = -\frac{\partial\phi(\omega,\tau)}{\partial\omega} \tag{2}$$

where $\phi(\omega,\tau)$ is the phase at at the point in the spectrogram with frequency $\omega$ at time $\tau$ and $\hat{\omega}(\omega,\tau)$ and $\hat{\tau}(\omega,\tau)$ are the reassignment operators that relate the amount of reassignment in frequency and time, respectively, at the corresponding time-frequency point. The frequency reassignment operator (1) can be considered a channelised version of the instantaneous frequency, while the temporal reassignment operator in (2) is the group delay [17].

The reassignment operators in (1) (2) can be estimated using different approaches. Perhaps the most well known of these is the reassignment method [8] which derives the frequency reassignment operator

$$\frac{\partial\phi(\omega,\tau)}{d\tau} = \Im\left(\frac{S_D(\omega,\tau)\times S^*(\omega,\tau)}{|S(\omega,\tau)|^2}\right) \tag{3}$$

where $S^*$ is the complex conjugate of a spectrogram, $S$, calculated using the STFT with a given window, $w$, and $S_D$ is a spectrogram for the same signal calculated using a derivative window $w_D = \frac{dw(t)}{dt}$, which can be calculated explicitly for many windows of interest. The temporal reassignment operator is given by as

$$\frac{\partial\phi(\omega,\tau)}{d\omega} = \Re\left(\frac{S_T(\omega,\tau)\times S^*(\omega,\tau)}{|S(\omega,\tau)|^2}\right) \tag{4}$$

where $S_T$ is a STFT calculated using a time derivative window, specified by $w_T(t) = w(t)\times t$. Other high resolution approaches include the derivative method [10] [9], which estimates the derivative of the signal rather than the window, and the difference method [11] which calculates the differences between time frequency bins. It has been shown previously that such methods are equivalent to each other under certain conditions [10] [9].

Higher order derivatives can be used to infer further information from signals. For example the frequency slope of linear chirped elements can be estimated using the second phase derivative with respect to time [9] [10]. The use of the second mixed derivative $\frac{\partial^2\phi(\omega,\tau)}{\partial\tau\partial\omega}$, which we refer to here as the group delay slope for simplicity, relates either the rate of change of the instantaneous frequency relative to the frequency, or the rate of change of the group delay relative to time. Similar to (1) (2) this can be estimated [17] through the use of windows

$$\frac{\partial^2\phi(\omega,\tau)}{\partial\tau\partial\omega} = -\frac{\partial\hat{\tau}(\omega,\tau)}{\partial\tau} = \frac{\partial\hat{\omega}(\omega,\tau)}{\partial\omega} - 1 = \Re\left(\frac{S_{TD}(\omega,\tau)S^*(\omega,\tau)}{|S(\omega,\tau)|^2}\right) - \Re\left(\frac{S_T(\omega,\tau)S_D(\omega,\tau)}{S^2(\omega,\tau)}\right) \tag{5}$$

where $S_{TD}$ is the the spectrogram calculated using the window $w_{TD} = t\times\frac{dw(t)}{dt}$ which applies the ramp function to the derivative window, $_D$. It is stated in [17] that for transients, $\frac{\partial^2\phi(\omega,\tau)}{\partial\tau\partial\omega}\approx 0$, while for stationary tonal elements $\frac{\partial^2\phi(\omega,\tau)}{\partial\tau\partial\omega}\approx -1$, as seen in Fig. 2. The use of this proposition is demonstrated for estimating parts of the spectrogram that relate to glottal pulses and harmonic elements of speech in [17], however it does not seem to have been previously exploited in music processing. We employ the mixed derivative (5) here as a part of the proposed onset detection system.

### 2.2. Onset detection

Onset detection is typically effected by flux methods [1] [5] that compare subsequent time frames in a spectrogram. The best known of these approaches is spectral flux which compares frames of a magnitude spectrogram $S\in R^{M\times N}$

$$SF(n) = H(|S(m,n)| - |S(m,n-1)|) \tag{6}$$

where $H()$ is the half wave rectifier function. Different variants on this approach have been employed, including different features, such as the phase difference [1] and different distance measures [2]. More advanced versions of spectral flux incorporate vibrato suppression methods [18] [5]. A good overview of such variants was originally given in [1] while a more recent compilation of features is given in [2]. Typically onset detection functions (ODF), such as *SF* (6) are non-negative and are post-processed using e.g. median filtering and thresholding in order to estimate the onsets [1].

A feature of particular interest here is the phase slope function, or equivalently the temporal reassignment operator (2), employed in [15] [16] which differs from most other functions in that it is not energy based, and naturally possesses negative and positive values. The authors of [15] describe how, unlike most ODFs, the zero crossings in the temporal reassignment feature $\frac{\partial\phi(\omega,\tau)}{\partial\omega}$ can describe onset locations. This is performed in multiple bands of the spectrogram, with temporal smoothing and a goodness function derived for each band, before summing to derive an onset detector function. These choices of post-processing are made in order to capture soft onsets. In the next section we build upon this approach, and propose modifications and additions specifically for the purpose of hard onset detection.

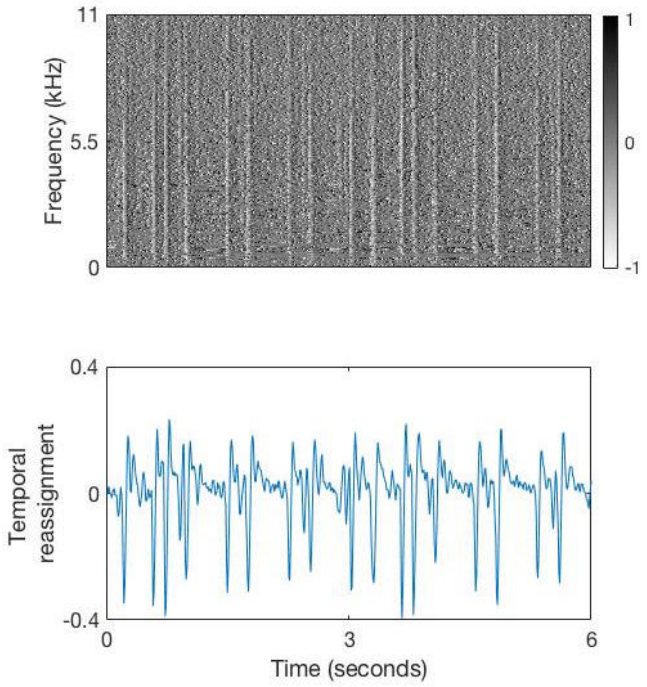


**Fig. 1**. Temporal reassignment spectrogram (top) and associated $GD$ feature (bottom)

## 3. PROPOSED APPROACH

We devise an onset detector that takes into account signal properties that may be associated with (semi-) percussive sounds. The onsets that are to be detected are generally of a transient nature and therefore are localised in time with a broadband spectrum, while the frequency spectrum away from onsets should be sparse with relatively few active tonal components. Following from this, it can be considered that many time-frequency points in the locality of a transient should be reassigned towards it, in a temporal sense. In this case the phase slope, or group delay, as used in [15] should provide a suitable feature, for the detection of such transient hard onsets. In order to capture a smooth feature, a relatively large of window of $\sim 92ms$ using 2048 samples at a sampling frequency of $22.05kHz$ is used, with a hop size of 220 samples, $\sim 10ms$, between frames. As the onset is considered broadband, the sum of the temporal reassignment function calculated using (4) is taken at each time frame:

$$GD(\tau) = \sum_{\omega=0}^{\bar{\omega}} \frac{\partial \phi(\omega, \tau)}{\partial \omega} \tag{7}$$

where $\bar{\omega}$ is a maximum frequency, rather than using a multiband approach such as in [15]. Care is taken to omit spurious values of the group delay, as may be found at low energy points of the spectrogram [17], by setting to zero elements

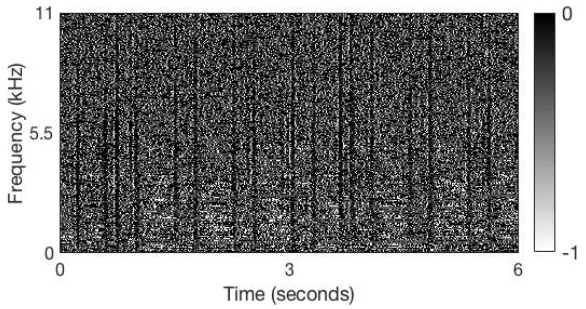


**Fig. 2**. Group delay slope spectrogram with vertical transients and horizontal tonal elements visible

that are reassigned outside the temporal support of a window, and temporal smoothing of $GD$ is performed by mean filtering over immediately adjacent time frames. An example of deriving $GD$ from the temporal reassignment feature is shown in Fig. 1. The feature $GD$ is then used to assign candidate onsets using positively sloped zero crossings in $GD$, similar to [15] which can be denoted by the set

$$Z = \{\tau | GD(\tau) < 0; GD(\tau + 1) > 0\}. \tag{8}$$

At each crossing, $\tau^i \in Z$, the height from the local minimum, $\tau^i_{min}$ to the local maximum $\tau^i_{max}$ is taken

$$\Delta_{\tau^i} = GD(\tau^i_{max}) - GD(\tau^i_{min}) \tag{9}$$

However, the cumulative slope weights $\Delta_\tau$, which possess no energy characteristic, are not found to be sufficient to derive a good onset detector, and a further weighting is applied. In this case the group delay slope (5) is employed to define a set of time-frequency points that are estimated to be transient in nature, at each $\tau^i \in Z$ which can be written as

$$\sigma_{\tau^i} = \{\omega | \frac{\partial^2 \phi(\omega, \tau^i)}{\partial \tau \partial \omega} > -0.2\}. \tag{10}$$

The final onset detection, W, then includes the magnitude of the spectrogram at points in $\sigma^i_\tau$

$$W_{\tau^i} = \sum_{\sigma_{\tau^i}} |S(\sigma_{\tau^i}, \tau^i)| \times \Delta_{\tau^i}. \tag{11}$$

Thresholding of W is subsequently performed in order to finally determine the onsets, as is typically performed in onset detection systems [1] [5].

## 4. EXPERIMENTS

Onset detection experiments were performed to assess the proposed approach, on solo piano and drum pieces. The superflux [5] and complexflux [14] were run for comparison, using the reference implementations freely available freely on the internet. The piano dataset used was the MAPS dataset

| | $P$ | $R$ | $\mathcal{F}$ |
|---|---|---|---|
| Superflux [5] | 89.3 | 86.7 | 88.0 |
| Complex [14] | 89.6 | 85.5 | 87.5 |
| Proposed | **96.0** | **90.2** | **93.0** |

**Table 1**. Onset detection results on the EnSTDkCL dataset

| | $P$ | $R$ | $\mathcal{F}$ |
|---|---|---|---|
| SuperFlux [5] | 79.6 | **70.1** | 74.6 |
| Complex [14] | 80.6 | 69.5 | 74.7 |
| Proposed | **81.8** | 70.0 | **75.4** |

**Table 3**. Onset detection results on ENST drums dataset

[19]. Similar to [4] we use the two different datasets of live recordings of an automatic Disklavier piano, which is generally considered to provide a reliable ground truth, although some have noted some minor alignment issues[20]. These two datasets are referred to as *EnSTDkCl* and *EnSTDkAm* which where recorded with the microphone close to the piano, and in the room ambience, respectively. 30 pieces were extracted from each dataset, with the first $30s$ of each piece used for the experiments. A small subset of the ENST drum dataset [21] is used, consisting of 27 pieces, randomly selected, with a total of 2293 onsets, with a mixture of pieces played with sticks, drums and rods.

Merging of onsets within $30ms$ of each other in the ground truth is performed, as seen in [4] [5]. For all approaches a sweep of a threshold parameter applied to the maximum value of $W$ is performed and results relating to the optimal threshold measure are given, similar to the approach in [22]. A sweep over an offset parameter spaced by $5ms$, as may account for the distance to the microphone, is also taken with optimal results given again. True positives, $tp$, are denoted when a ground truth onset is found within $50ms$ of an estimated onset, while care is taken to assign each estimated onset to only one ground truth onset, and vice versa. False positives, $fp$, and false negatives, $fn$, are also denoted, from which the standard Precision (P), Recall (R) and $\mathcal{F}$-measure metrics [22] are derived, which are recorded in percentage scores.

The results for the piano datasets are seen in Tables 1 & 2 for the close and ambient recordings respectively. It is seen in the tables that there is little difference between complexflux and superflux for this task, as also recorded in [4], which is to be expected perhaps, as the signals do not consist of vibrato elements. In both cases, improvements are seen using the proposed method, around $5\%$ on the close recording and $10\%$ for the ambient recording. Further analysis shows that the proposed method achieves $\mathcal{F}$-measures of $86.3\%$ and $83.9\%$ for the close and ambient recording environments when a smaller tolerance window of $25ms$ is used. In particular, these results for the ambient recordings are better than those of the flux methods using the standard $50ms$ tolerance. It is difficult to determine whether the difference in the improvements with the proposed method, relative to the datasets, is due to its ability to deal with the ambient environment, or it may be that the high results for the flux methods in the case of the close recordings leave little room for improvement. Conversely, little difference is in seen the results for the drums dataset, as can be seen in Table 3, where the proposed method is seen to perform only slightly better than the previous approaches. It was observed, in these experiments, that the $\mathcal{F}$-measures for the pieces using drumsticks were higher than those on the piano datasets, while results on pieces using rods or brushes with less localised onsets were lower.

| | $P$ | $R$ | $\mathcal{F}$ |
|---|---|---|---|
| Superflux [5] | 83.3 | 78.3 | 80.7 |
| Complex [14] | 81.2 | 79.4 | 80.3 |
| Proposed | **95.0** | **86.5** | **90.6** |

**Table 2**. Onset detection results on the EnSTDkAm dataset

It would seem that the proposed approach is useful in the specific scenarios described. Although not reported here, other variants incorporating the group delay slope were experimented with, such as a version of spectral flux with masking based upon the group delay slope, and a version in which the weighting, $\Delta_{\{}tau$, was not employed. However, none of these other variants seemed to improve upon the flux methods used for comparison. It would seem that using the (group delay slope masked) energy at a given point is best. Possibly this is due to its effect of penalising lower energy reverberation that may be detected as candidate onsets, which may have a similar shape in terms of the group delay feature.

## 5. CONCLUSIONS

We have proposed a novel onset detector for the purpose of hard onset detection. Experimental validation shows that, for such a purpose, the proposed approach outperforms the state-of-the-art general purpose onset detection methods. Meanwhile the usefulness of the group delay slope has been demonstrated. We believe that the proposed system will extend to other semi-percussive instruments. Future work will include experimenting with other types of data, exploring performance in presence of other sounds. Different window sizes, or possibly multi-scale approaches, might be explored in order to try to enhance the temporal tolerance of the approach, while other high resolution approaches, such as the derivative and difference methods should also be compared for this type of application. Such an onset detector may also be used in a piano transcription system such as [4].

## 6. REFERENCES


[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, Sept 2005.

[2] "Multi-feature beat tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 816–825, April 2014.

[3] E. Quinton, K. O'Hanlon, S. Dixon, and Mark B. Sandler, "Tracking metrical structure changes with sparse-NMF," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[4] J. J. Valero-Mas, E. Benetos, and J. M. Inesta, "Assessing the relevance of onset information for note tracking in piano music transcription," in *2017 AES International Conference on Semantic Audio*, 2017.

[5] S. Bock and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proceedings of the 16 International Conference on Digital Audio Effects (DaFX)*, 2013.

[6] S. Hargreaves, A. Klapuri, and M. Sandler, "Structural segmentation of multitrack audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2637–2647, Dec 2012.

[7] F. Eyben, S. Bock, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long short-term memory neural networks," in *Proceedings of 11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[8] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, May 1995.

[9] B.Hamilton, P. Depalle, and S. Marchand, "Theoretical and practical comparisons of the reassignment method and the derivative method for the estimation of the frequency slope.," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 345–348.

[10] X. Wen and M. B. Sandler, "Notes on model-based non-stationary sinusoid estimation methods using derivatives," in *Proceedings of the 12th Int. Conference on Digital Audio Effects*, 2009.

[11] S. Marchand, "The simplest analysis method for non-stationary sinusoidal modeling," in *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx), 2012*, 2012.

[12] M. Khadkevich and M. Omologo, "Time-frequency reassigned features for automatic chord recognition," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 181–184.

[13] P. Smaragdis and M. Kim, "Non-negative matrix factorization for irregularly-spaced transforms," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[14] S. Bock and G. Widmer, "Local group delay based vibrato and tremolo suppression for onset detection," in *Proceedings of ISMIR - International Conference on Music Information Retrieval (ISMIR)*, 2013.

[15] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in *Proceedings of ISMIR - International Conference on Music Information Retrieval*, 2008, pp. 653–658.

[16] A. Holzapfel, Y. Stylianou, A. C. Gedik, and B. Bozkurt, "Three dimensions of pitched instrument onset detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1517–1527, 2010.

[17] K. R. Fitz and S. A. Fulop, "A unified theory of time-frequency reassignment," *CoRR*, vol. abs/0903.3080, December 2009.

[18] N. Collins, "Using a pitch detector for onset detection," in *International Conference on Music Information Retrieval (ISMIR)*, 2005.

[19] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, August 2010.

[20] A. Cogliati, Z. Duan, and B. Wohlberg, "Context-dependent piano music transcription with convolutional sparse coding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2218–2230, Dec 2016.

[21] "Enst-drums: an extensive audio-visual database for drum signals processing," in *Proceedings of ISMIR - International Conference on Music Information Retrieval (ISMIR)*, 2006.

[22] K. O'Hanlon and M. B. Sandler, "An iterative hard thresholding approach to $\ell_0$ sparse hellinger nmf," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4737–4741.