# Onset Detection via Separation of Harmonic Content from Musical Notes

Alejandro Delgado Castro
*Department of Electronic Engineering*
*University of York*
York, United Kingdom
adc533@york.ac.uk

Giorgos Siamantas
*Independent Researcher*
Hamburg, Germany
g.siamantas@gmail.com

John E. Szymanski
*Department of Electronic Engineering*
*University of York*
York, United Kingdom
john.szymanski@york.ac.uk

***Abstract*—A novel method for onset detection in single-channel audio recordings is presented and evaluated. Here, source separation techniques are used as a preprocessing stage for extracting the harmonic content of musical notes. The residual channel is then used to estimate an onset detection function whose peaks align with note transitions. Several tests are conducted on a selected dataset in order to evaluate its performance and compare with alternative algorithms. The results provide evidence that the proposed residual-based method can achieve comparable levels of accuracy without the need of previous training stages.**

***Index Terms*—onset detection, source separation, spectral filtering, multipitch estimation, music information retrieval.**

## I. Introduction

Within the field of Music Information Retrieval (MIR), the detection of onsets in music has a significant importance in obtaining further high-level features, such as beat or rhythm, and it is also essential in Automatic Music Transcription (AMT) [1].

Onset detection refers to the task of finding the time instant at which any particular note event starts. However, there are several factors that can make this a difficult task, for example the number of sources playing simultaneously, their relative volumes, and how close they are in time.

Algorithms designed to automatically detect onsets in musical recordings usually require the estimation of an Onset Detection Function (ODF), which is derived from the input data and whose peaks normally correspond to particular changes in the signal. Obtaining the ODF usually implies breaking the audio signal into a set of consecutive overlapping frames, and then the evolution of some particular feature within the new representation is observed across time [2]–[5].

More recent approaches have incorporated techniques from machine learning and support vector machines. In this case, the input data is also a time-frequency representation of the signal, and neural networks are commonly used to obtain the final position of the onsets [6].

In this study, source separation techniques based on pitch tracking and spectral filtering, are applied to extract the harmonic content of notes in order to obtain a residual channel, in which the position of transients should be clearer.

This work was supported by the University of Costa Rica (UCR) and the Costa Rican Ministry of Science, Technology and Telecommunications (MICITT).

The remainder of this paper is structured as follows. In Section 2 a review of previous onset detectors is presented, focussing on the method used to obtain the ODF. Section 3 describes the harmonic-content separation process and the residual channel generation. The proposed onset detection algorithm is presented in Section 4 and its evaluation is presented in Section 5. Finally, Section 6 concludes the paper and gives some ideas for further work.

## II. Onset Detection Algorithms

According to Shao et. al. [1] probably the most difficult challenge for an onset detector is to obtain an ODF based on a chosen type of transformation followed by feature extraction. This function has to detect relevant changes in the audio signal whilst remaining robust enough to reject any possible interference created by intrinsic playing styles, e.g. vibrato, glissandi and ornamentation, or any form of impulsive noise.

Classical onset detection strategies are usually classified either as energy-based or phase-based algorithms. In the first case, onsets are detected by looking at sudden changes in the energy of the signal, whilst methods in the second group assume that any change in instantaneous frequency could be an indicator of a possible onset. Although these two initial approaches proved to be useful, energy-based detectors cannot distinguish between increases and decreases in amplitude of the signal, while phase-based systems are susceptible to noise introduced by components with no significant energy [3].

A major improvement to onset detection was proposed by André Holzapfel et. al. in 2010 [5]. In this study, three separated ODF's were estimated according to the phase slope, spectral flux, and fundamental frequency contours, estimated by using the YIN algorithm. Detections coming from these three functions were added and smoothed to generate a strength signal from which the final location of the onsets were selected by peak-picking. Tests were conducted using a specific dataset of pitched instruments, considering evaluation aspects defined by the Music Information Retrieval Evaluation Exchange (MIREX). The authors reported good levels of accuracy for isolated instruments, while lower levels were observed for complex mixtures.

More recently, the use of sparse decomposition techniques has been explored in onset detection systems. Shao et. al. [1]

sparsely decomposed the input signal using Matching Pursuit (MP) and the resulting time-frequency representation was used as the input for an hybrid detection algorithm combining the Degree of Explanation (DE) and the Change of Partials (CP). The combined results showed improved detection accuracy on a dataset comprising 2050 onsets.

Stasiak and Mońko proposed an algorithm for onset detection based on Convolutional Neural Networks (CNN) [6]. In this case, the network was fed with spectrogram fragments in the form of images with 15 columns and 80 rows. The output of the network was treated as a classical ODF and a fixed threshold was used to detect peaks. The dataset for testing was similar to the one used in [5] and a subset was utilized for training purposes.

The influence of noise in onset detection was recently studied by Maka in [7]. His research aimed to evaluate the performance of five well-established algorithm under noisy conditions. Four types of background noises were selected for the tests, namely cars, shop noises, rain and wind sounds, and the interference created by the rain was found to be the most significant factor in terms of reducing the final accuracy during onset detection.

## III. Separation of Harmonic Content from Musical Notes

The previous section mentioned important advances that have been proposed in onset detection. This initial study aims to enhance the visibility of onsets by first identifying and extracting some of the masking harmonic content that is present in the input signal. In this section we present the separation algorithm that is used to extract the harmonic content from musical notes, which constitutes the preprocessing stage of the proposed system.

### A. Time-Frequency Representation

The standard Short-Time Fourier Transform (STFT) was selected as a basis to represent signals. Although other time-frequency representations have been studied as well, for example the Correlogram in [6] and the Gammatone Filter Bank in [7], results show that the classic spectrogram still represents the best choice, due to the low number of artifacts that it generates. To obtain a good time resolution, which is an essential aspect in onset detection, a frame size of 2048 samples is considered, with 75% overlap. A Hanning window function has been selected to smooth out edge effects in each frame.

### B. Estimation of the Pitch Trajectory

Profiling the pitch trajectory of a harmonic sound is necessary to identify the location of its partials. Depending on how many sources are present, and the degree of overlap between them, the task of estimating a reliable trajectory for each of them is highly challenging on its own. Some common problems in pitch estimation include octave errors, misleading detection of silence gaps, and incomplete note tracking.

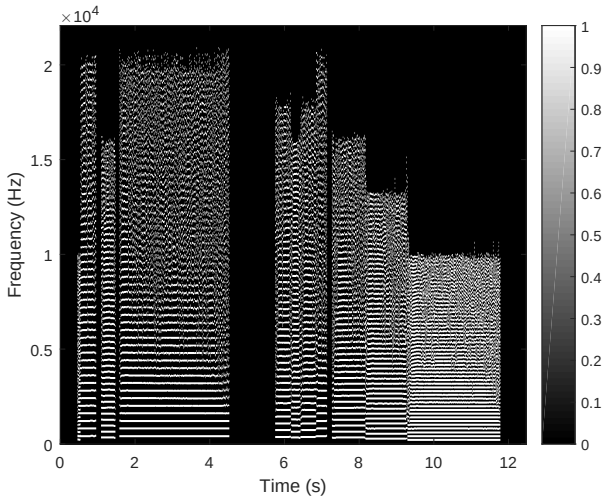


Fig. 1. *Spectral mask used to extract the harmonic content from a set of notes played by a saxophone.*

In this study, a modified version of the multipitch estimation algorithm, developed by Duan et. al. [8], is used to obtain a framewise fundamental frequency estimate for the dominant harmonic source present in the input signal. The method was selected considering its overall good performance and flexibility. An implementation of the algorithm is available on the author's website [9].

### C. Spectral Filter and Residual Channel

The idea of designing spectral filters to separate harmonic content from audio recordings was proposed by Every and Szymanski in [10], and further explored by Siamantas in [11]. Pitch information is used to design an extraction mask for a set of harmonic partials associated to one particular source. The method incorporates some degree of flexibility to allow small deviations between the real and ideal position of harmonic partials.

The mask can be then constructed by setting to one those time-frequency tiles of the mask that correspond to the frequency bins where the harmonic partials are located, while setting the rest of the values to zero. Figure 1 shows a spectral mask used to extract the harmonic content from a set of saxophone notes.

If $M_H(t, f)$ denotes the spectral mask in the time-frequency plane for the harmonic content, the mask to obtain the residual can be obtained by means of the following relation.

$$M_R(t, f) = 1 - M_H(t, f) \tag{1}$$

The residual signal is generated by multiplication between the original magnitude spectrogram and the residual mask, followed by inverse transformation in which the phase information of the original signal is used to complete the reconstruction.

## IV. Residual-Based Onset Detector

Very little work has looked at the residual signal resulting from a separation process as a means of obtaining useful information [11]. In general, after the harmonic content of every source has been extracted, the residual channel can be expressed in the following way.

$$x_{res}(t) = x(t) - \sum_{j=1}^{J} \hat{s}_j(t) \tag{2}$$

Where $x(t)$ is the original time domain signal, $\hat{s}_j(t)$ is the estimated *j*-th harmonic source, and $J$ is the total number of harmonic sources present. The following cases can be distinguished regarding the content of $x_{res}(t)$, mentioned in order of importance for the present approach.

- Harmonic content as a result of misleading pitch estimation.
- Non-harmonic content of a fast and impulsive nature, for example onset transients of harmonic notes and non-pitched percussive sources.
- Non-harmonic content of a structured broadband-noise nature, for example breathiness in wind instruments.

If harmonicity or near-harmonicity is used as the primary part of a source model and the input signal consists of discrete musical note events, the signal content that ends up in the residual channel can provide access to valuable timing information. The following subsections describe the proposed method.

### A. Onset Detection Function

The residual channel obtained after extracting the harmonic content from a set of saxophone notes is presented along with the original signal in Figure 2. In this graph it is evident that the residual presents low energy levels during silence gaps or during the sustain of each note, whilst short bursts of energy appear during note transitions. Two excessively long offsets can also be observed in the residual, starting at $t = 4.5$s and $t = 11.8$s. They are due to errors in the pitch trajectory since Duan's algorithm was unable to follow the whole duration of these notes.

As these energy bursts align with note transitions, the residual channel seems to be a reasonably good basis for an ODF. In order to obtain a stable detection function though, some specific transformations have to be carried out. The first step is to half-wave rectify the residual signal before computing its upper root-mean-squared envelope curve using a window size of 1000 samples.

The resulting envelope curve is then downsampled by a factor of 140 and the mode of the resulting curve is then subtracted. Finally, the envelope curve is smoothed by means of a moving average filter of length 6. The result is an ODF that emphasizes the peaks at the transitions between musical notes. Figure 3 summarizes the way in which the ODF is obtained from the residual channel.

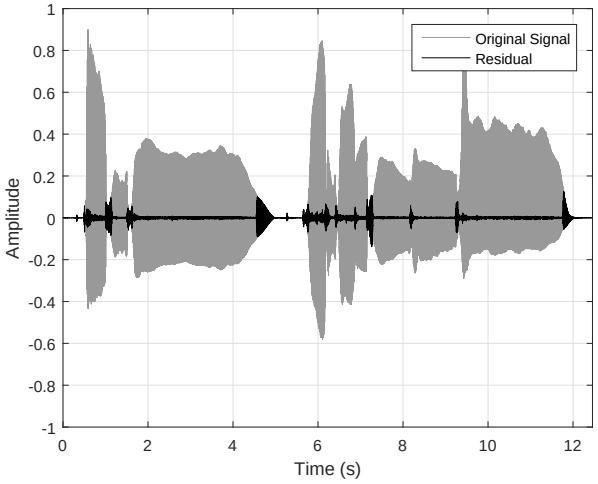


Fig. 2. *Residual channel (black) presented on top of the original signal (grey).*

### B. Threshold and Peak-Picking

Given the Onset Detection Function ODF$(n)$, the next step is to identify all peaks with a prominence higher than a specified *threshold*. There are two different approaches to define the threshold, according to Bello et. al. [2]: *fixed* and *adaptive*. The proposed algorithm uses a combination of both. The reference threshold, denoted by $T$, is defined as follows.

$$T = \mathrm{E}[\mathrm{ODF}(n)] \tag{3}$$

Instead of using the value of $T$ as a minimum peak height (fixed threshold), the value is used as minimum peak prominence. That is, each detected peak must have a vertical drop of more than $T$ units from the peak on both sides, without encountering either the end of the function or a larger intervening peak. This feature gives some level of adaptability to the threshold and helps to detect real low-magnitude peaks without picking a significant number of noisy ones.

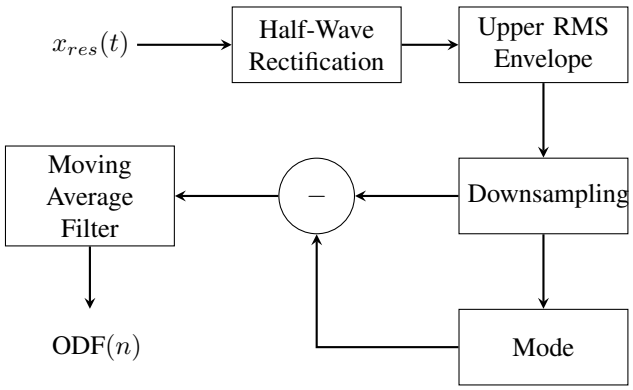


Fig. 3. *Generation of the Onset Detection Function (ODF) from the residual channel* $x_{res}(t)$.

### C. Peak Pairs

Considering that the energy within the residual channel is mostly related with offsets and onsets, the occurrence of peak-pairs is common. They are pairs of detected peaks corresponding to the same note transition, the first one corresponding to an offset and the second one to an onset.

To detect peak pairs, the minimum distance $S_M$ has to be calculated based on the distance between all detected peaks in $ODF(n)$. Denoting the set of distances between adjacent peaks as $P_{sep}$, then the value of $S_M$ is estimated by the following relation.

$$S_M = \frac{\text{median}(P_{sep})}{2} \tag{4}$$

Two consecutive peaks in $ODF(n)$, whose distance is equal or smaller than $S_M$, are considered as a peak pair and therefore, treated as a single onset.

### D. Onset Selection

Due to the nature of the proposed ODF, peaks tend to appear later than the proper onsets. Hence, their actual position has to be estimated in accordance with the location of its corresponding peak.

Considering the *i*-th detected peak, centred at sample $m_i$, then the onset position, denoted $O_{P_i}$, is found within the interval of samples $[m_i - \delta, m_i]$, according to the following equation.

$$O_{P_i} = \underset{p \in [m_i - \delta, m_i]}{\text{argmin}} \ \text{ODF}(p) \tag{5}$$

Where the value of $\delta$ has been empirically selected as 15 samples for single detected peaks and 30 for peak pairs.

Spurious peaks related with long offsets can also be rejected by taking advantage of their slower decay and asymmetry. The asymmetry can be estimated by observing the number of samples between the peak position and the position at which the function crosses the threshold $T$. If this value is greater than the distance between the peak and its onset position, the peak and its onset position are rejected. Figure 4 presents the estimated onset positions for the notes played by a saxophone.

## V. Evaluation

In this section the proposed onset detection algorithm is evaluated with regards to its performance on several audio signals. The selected test recordings and the metrics used for evaluation purposes are described in the following subsections.

### A. Dataset

Evaluation of performance was conducted using a set of twenty three monophonic audio excerpts, sampled at 44.1 kHz, taken from the same database used by Holzapfel et. al. in [5]. It contains wind and string musical instruments like violin, piano, guitar, clarinet and trumpet. A total number of 726 real onsets are present, with annotations provided by André Holzapfel, which were further revised by Sebastian Böck.

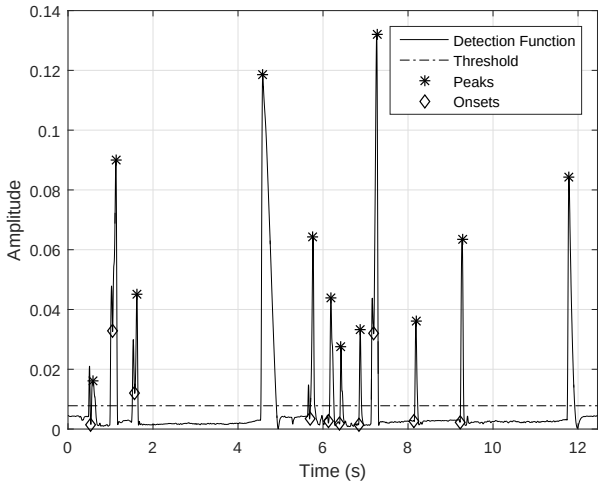


Fig. 4. *Final onset positions for a set of notes played by a saxophone. The final set of accepted peaks are marked with asterisks whilst diamonds indicate onset positions.*

### B. Methodology

The performance of the algorithm was evaluated using the selected dataset and annotations to compute the F-Measures, in the same way as it was used in the MIREX onset detection evaluation. A tolerance window of $\pm 50$ms was considered, and penalties related to double detections and merged onsets were also incorporated. Two well-known onset detection systems and one state-of-the-art algorithm were selected for comparison and applied to the same input signals. These systems are briefly described below.

- **AUBIO:** Well-known open-source algorithm designed by Paul Brossier [12].
- **QMARY:** Well-known complex domain-based method developed at Queen Mary College [4].
- **CNN:** State-of-the-art onset detector based on a trained CNN developed by Stasiak and Mońko [6].

### C. Results and Discussion

Table I shows the results obtained by the proposed method and the alternative systems when applied to the selected dataset. In each case, the F-measure is given as a percentage.

TABLE I
*F-Measures obtained using the proposed algorithm and alternative onset detectors.*

| Algorithm | Type of Instrument | | Average |
|---|---|---|---|
| | String | Wind | |
| PROPOSED | 54.7% | 61.6% | 58.6% |
| AUBIO | 55.4% | 63.5% | 60.0% |
| QMARY | 67.9% | 70.6% | 69.5% |
| CNN | 86.1% | 88.6% | 87.5% |

The first point to note it that the state-of-the-art algorithm is the one with the highest accuracy in onset detection. This result is not surprising since the CNN algorithm creates learned models of the input signals during a previous training stage,

which can also be significantly hard and requires many audio examples. The proposed method, on the other hand, runs without any previous knowledge of the input recordings and surprisingly, its performance is similar to that exhibited by alternative onset detection methods which do not incorporate learning features either.

It was also observed that Queen Mary's onset detector was the one generating the highest number of false positives. If the number of detections is high, the probability of detecting many real onsets increases, which leads to a higher performance in terms of the F-Measure. The reason for this lies in the actual definition of the F-Measure, which penalises more false negatives compared with false positives. Considering the proposed algorithm and the nature of its ODF, the number of detections happened to be lower or equal to the number of real onsets, for most of the audio excerpts, which influenced the final F-Measure value negatively.

Finding a proper onset time for each selected peak in the ODF proved to be a difficult task with significant impact on the overall performance of the system. Several cases were observed in which the local minimum, before a real onset peak, occurred outside the window of 50 ms defined by MIREX and therefore, each of these detections were counted as one false positive and one false negative.

It is also worth mentioning that the quality of the residual channel, on which the proposed ODF is based, depends on the accuracy of the separation of the harmonic content, and hence, on the quality of the estimated pitch trajectory. Although the pitch tracking stage used here showed good results in detecting the fundamental frequency in most of the notes, it also failed in following their complete duration, which created significant peaks at the offsets. In those cases where an offset was far apart from the following onset, the symmetry of the peak was used to differentiate the onset peak from the offset one, but this assumption was not sufficient to handle all possible cases. Therefore, extracting high quality pitch information from the input signal is an important aspect that will be explored in future research.

Finally, considering the type of instrument, it can be observed that all considered onset detection methods showed slightly higher accuracy for wind instruments than for string ones. However, the average difference on F-Measure is less than 5%, which is too small to conclude that the type of instrument influenced the accuracy of the onset detector. Considering that the score was different in every audio excerpt, these differences in accuracy can also be attributed to each particular performer, who chose a different playing style and tempo in each case.

## VI. Conclusions and Further Work

In this paper a novel residual-based onset detection system was introduced, in which the separation of harmonic content from musical notes is used as a preprocessing stage. The proposed ODF and onset selection strategy exhibit a comparable performance, in the sense of F-measure, when applied to a dataset comprising several pitched instruments. No information is required *a priori* for the system operation.

The energy left in the residual channel, after spectral filtering, was shown to constitute an interesting basis for an onset detector. Since the spectral filter is directly guided by an automatic pitch detector, no signal dependent parameters or previous training stages are required.

In future work, the possibility of improving the automatic pitch estimation process will be explored, in order to generate more accurate pitch trajectories. These improvements will significantly help in achieving higher quality residuals and detection functions.

The ODF interpretation can also be refined to obtain a sharper selection of real peaks, decreasing the number of missed onsets. Finally, a more sophisticated strategy can also be introduce to locate more precisely the onset times, once their corresponding peaks have been identified.

## Acknowledgement

The authors would like to thank the University of Costa Rica (UCR), and the Costa Rican Ministry of Science, Technology and Telecommunications (MICITT) for their support in funding this research. Also, special thanks to Barłomiej Stasiak and André Holzapfel for their valuable cooperation which significantly enriched this paper.

## References

[1] X. Shao, W. Gui, and C. Xu, "Note Onset Detection Based on Sparse Decomposition," *Multimedia Tools and Applications*, no. 172, pp. 2613–2631, 2015.

[2] 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. 1035–1046, 2005.

[3] S. Dixon, "Onset Detection Revisited," *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06)*, pp. 1–6, 2006.

[4] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "Complex Domain Onset Detection for Musical Signals," *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03)*, no. 1, pp. 6–9, 2003.

[5] 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.

[6] B. Stasiak and J. Mońko, "Analysis of Time-Frequency Representations for Musical Onset Detection with Convolutional Neural Network," *Proceedings of the Federated Conference on Computer Science and Information Systems*, vol. 8, pp. 147–152, 2016.

[7] T. Maka, "A Comparative Study of Onset Detection Methods in the Presence of Background Noise," *2016 International Conference on Signals and Electronic Systems (ICSES)*, pp. 51–56, 2016.

[8] Z. Duan, B. Pardo, and C. Zhang, "Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks and Non-Peak Regions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 8, pp. 2121–2133, 2010.

[9] Z. Duan, "Multi-Pitch Analysis," http://www.ece.rochester.edu/~zduan/multipitch/multipitch.html.

[10] M. R. Every and J. E. Szymanski, "Separation of Synchronous Pitched Notes by Spectral Filtering of Harmonics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1845–1856, 2006.

[11] G. Siamantas, "An Iterative, Residual-Based Approach to Unsupervised Musical Source Separation in Single-Channel Mixtures." PhD, University of York, 2009.

[12] P. Brossier, "AUBIO: Extraction of Annotations from Audio Signals," https://aubio.org/vamp-aubio-plugins/.