# EXTRACTING NOTE ONSETS FROM MUSICAL RECORDINGS

*Miguel Alonso*[†]*, Gaël Richard and Bertrand David*

GET–Télécom Paris
46 rue Barrault, 75634 Paris Cedex 14, France
{malonso,grichard,bedavid}@tsi.enst.fr

**ABSTRACT**

Automatic temporal segmentation of music signals into note onsets is central for a large number of audio applications. In this paper, we present a variation of a previously existing note onset detection method, based on the so-called spectral energy flux. The proposed algorithm has a lower computational cost and incorporates a more accurate estimation of the frequency content derivative, yielding better results for a wide range of music signals. The performance of the system was validated using a database of musical recordings containing 670 note onsets. This database was hand-labeled and cross validated by three annotators. Comparisons to previous work are also presented along with possible directions of future research.

**Keywords**: onset detection, differentiator filter, adaptive thresholding.

## 1. INTRODUCTION

Computer assisted music analysis is an increasingly active research area. In this field, automatic temporal segmentation of a music stream into note onsets plays an important role for numerous applications. For example, automatic transcription, rhythm parsing, music retrieval, audio editing and special effects.

The term musical note onset, perhaps better defined as a *phenomenal accent* [1], refers to discrete temporal events in an audio stream where there is a marked change in any of the perceived psychoacoustical properties of sound, i.e., loudness, timbre, and pitch. Throughout the present document, we adopt a signal processing approach and we strive to detect magnitude changes, harmonic changes and pitch leaps. That is, acoustic effects that can be heard and are musically relevant for the listener.

Robust onset detection for a wide range of music signals has proven to be a difficult task due to the large variety of instruments that can be employed, whether they play simultaneously or not and the different kinds of attacks and dynamic ranges that they can produce. In recent years a large effort has been invested to this problem. In [2] Bello provides an in-depth survey of the most commonly used methods. In general, these approaches can be divided into two main categories according to the working principle:

- *deterministic techniques*, they use time–frequency or time–scale features of the audio signal,
- *statistic techniques*, lie on the assumption that the signal can be described by a probabilistic model.

We propose in this article a model that lies into the first category. This system significantly improves the calculation of a method so-called the Spectral Energy Flux (SEF) or Spectral Difference (SD) by performing an accurate estimation of the derivative of the signal frequency content with respect to time at a relatively low computational cost.

After presenting in section 2 the algorithm description, we study its performance in section 3 and validate it on a 670 onsets database of hand-labeled music segments. Comparisons between our approach and the previous algorithm are also presented. Results are given as tables and ROC curves and they are followed by a discussion. Finally, section 4 highlights concluding comments and possible directions for future research.

## 2. METHOD DESCRIPTION

Early endeavors to detect note onsets in music signals used to process the amplitude envelope of the waveform as a whole. This approach has proved to be very vulnerable since note onsets can be easily masked in the bulk signal by continuous tones of higher amplitude. More recent advances [3] have shown that musical events are more likely detected after separating the music signal in frequency channels. There exists no consensus on an optimal frequency decomposition for onset detection, since several decompositions reported in the literature have led to comparable results.

An overview of the system that we propose in this article is presented in Figure 1. It uses the band-wise processing principle as motivated above. First, the signal is decomposed into frequency channels using the STFT (Short Time Fourier Transform). Each frequency band is processed as depicted in Figure 2 to find the time-location and intensity of its onset component. Then, contributions from all frequencies are summed, smoothed and thresholded. Finally, the system output, called the *detection function*, is a signal that bears peaks with magnitude and location related to the onsets intensity and position.

### 2.1. Spectral energy flux

The system that we present resides on the general assumption that the appearing of a phenomenal accent in an audio stream leads to a variation in the signal's frequency content at certain frequency components. For example, in the case of a violin producing pitched notes, the resulting signal will have a strong fundamental frequency that leaps in time and related harmonic components at integer multiples of the fundamental attenuating as frequency increases. In the

†Research supported by a grant from the Mexican Council for Science and Technology (CONACyT).

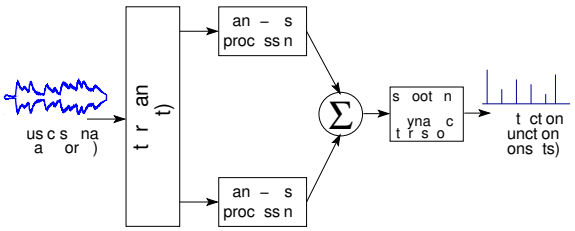


**Fig. 1**. Onset detection system overview.

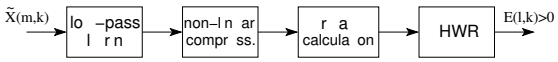


**Fig. 2**. Processing performed at each frequency channel.

case of a percussive instrument playing, the resulting signal will tend to have sharp energy boosts in the passband. The input audio signal is analyzed using the STFT, leading to

$$\tilde{X}(m,k) = \sum_{n=-\infty}^{\infty} w(Mm-n)x(n)e^{-j\frac{2\pi}{N}kn} \tag{1}$$

where $x(n)$ denotes the audio signal, $w(n)$ a finite-length sliding window, $M$ the hop size, $m$ the time (frame) index and $k = 0,\ldots,N-1$ the frequency (bin) index. To detect the above mentioned variations in the frequency content of the audio signal, previous approaches (for example [3,4]) have proposed the calculation of the derivative of $\tilde{X}(m,k)$ with respect to time

$$E(l,k) = \sum_{m} h(l-m)\,G(m,k) \tag{2}$$

where $E(l,k)$ is known as the Spectral Energy Flux (SEF) and where $h(m)$ is an approximation to an ideal differentiator

$$H(e^{j2\pi f}) \simeq j2\pi f \tag{3}$$

and where

$$G(m,k) = \mathcal{F}\{|\tilde{X}(m,k)|\} \tag{4}$$

is a transformation that accentuates the psychoacoustically relevant properties of $\tilde{X}(m,k)$.

In solving many physical problems by means of numerical methods, it is a challenge to seek derivatives of functions given in discrete points. For example, in [3, 4] authors propose a first order difference with $h = [1, -1]$, which is a rough approximation to (3). In this paper, we use a differentiator filter $h(m)$ of order $2L$ based on the formulae for central differentiation developed by Dvornikov in [5] which provides a much closer approximation to (3). The underlying principle is the calculation of an interpolating polynomial of order $2L$ passing through $2L+1$ discrete points which is used to find the derivative approximation. A comprehensive description of the method and its accuracy to approximate (3) can be found in [5]. The analytical expression to compute the first $L$ coefficients of an antisymmetric FIR differentiator is given by

$$g(i) = \frac{1}{i\alpha(i)} \tag{5}$$

with

$$\alpha(i) = \prod_{\substack{j=1 \\ j\neq i}}^{L} \left(1 - \frac{i^2}{j^2}\right) \tag{6}$$

and $i = 1,\ldots,L$. The coefficients of $h(m)$ are given by

$$h = [-g(L),\ldots,0,\ldots,g(L)]. \tag{7}$$

In our proposal, the transformation $G(m,k)$ calculates a perceptually plausible power envelope for frequency channel $k$ and is formed of two steps. First, psychoacoustic research on computational models of mechanical to neural transduction [6] shows that the auditory nerve adaptation response following a sudden stimulus change can be characterized as the sum of two exponential decay functions:

$$s(m) = \alpha e^{-m/T_1} + \beta e^{-m/T_2} \quad \text{for} \quad m \geq 0 \tag{8}$$

formed by a rapid decline component with time constant ($T_1$) in the order of 10 ms and a slower short-term decline with a time constant ($T_2$) in the region of 70 ms. This adaptation function performs energy integration, emphasizing the most recent stimulus but masking rapid modulations. From a signal processing standpoint, this can be viewed as two smoothing low-pass filters having a discontinuity that preserves edge sharpness and avoids dulling signal attacks. In practice, the smoothing window is implemented as a 2nd-order IIR filter with $z$-transform

$$S(z) = \frac{\alpha+\beta-(\alpha e^{-1/T_2}+\beta e^{-1/T_1})z^{-1}}{1-(e^{-1/T_1}+e^{-1/T_2})z^{-1}+e^{-(1/T_1+1/T_2)}z^{-2}}. \tag{9}$$

Figure 3 shows the role of the energy integration function after convolving it with a pitched channel of a signal's time–frequency representation.

The second part of the envelope extraction consists in a logarithmic compression. This operation has also a perceptual relevance since the logarithmic difference function gives the amount of change in a signal's intensity in relation to its level, that is

$$\frac{d}{dt}\log I(t) = \frac{\Delta I(t)}{I(t)}. \tag{10}$$

This means that the same amount of increase is more prominent in a quite signal [3, 7].

In practice, the algorithm implementation is straightforward, and is carried out as presented in Figure 1. The time–frequency representation (1) is computed using an $N$ point FFT. The absolute value of every frequency channel, $|\tilde{X}(m,k)|$ is convolved with $s(n)$. Since [6] only provides an expression to calculate the parameters of (8) from experimental data, these values were fixed to maximize the low-pass *vs.* time-spread trade-off. The smoothing operation is followed by a logarithmic compression. The resulting $G(m,k)$ is given by

$$G(m,k) = \log_{10}\left(\sum_{i}|\tilde{X}(i,k)|s(m-i)\right). \tag{11}$$

At those time instants where the frequency content of $x(n)$ changes and new frequency components appear, $E(l,k)$ exhibits positive peaks whose amplitude is proportional to the energy and rate of change of the new components. In a similar way, when frequency components disappear from $x(n)$, the SEF exhibits negative peaks, marking the *offset* of a musical event. Since we are only

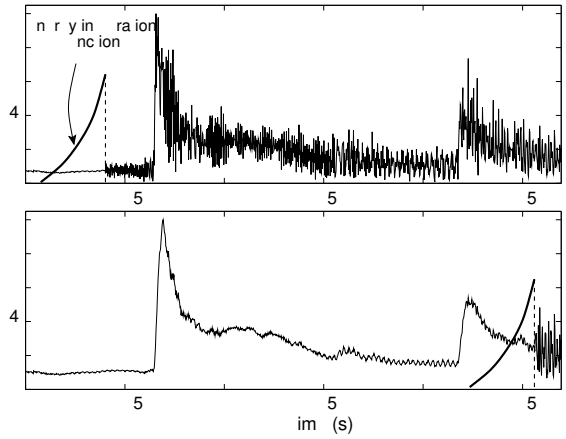


**Fig. 3**. The smoothing effect of the energy integration function emphasizes signal attacks but masks rapid modulations.

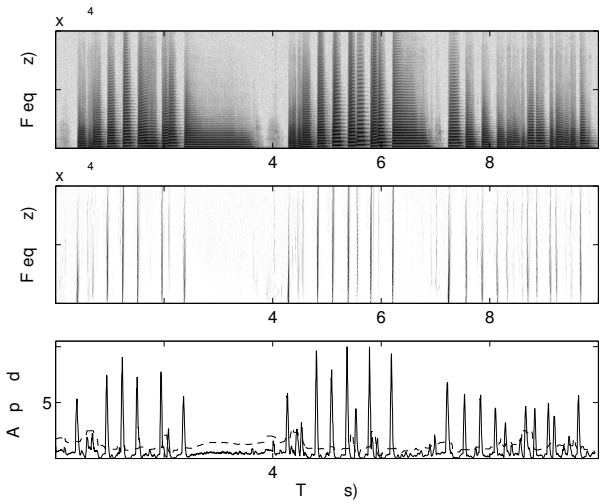


**Fig. 4**. From top to bottom: time–frequency representation for a trumpet signal, the corresponding SEF and the detection function with its respective adaptive threshold. The SEF image displays sharp vertical edges at those instants where the frequency content of the signal exhibits rapid variations.

interested in onsets, we apply a half-wave rectification (HWR) to $E(l,k)$, i.e., only positive values are taken into account. To find a global stationarity profile $v(l)$, also called *detection function*, contributions from all channels are integrated across frequency

$$v(l) = \sum_{\substack{k \\ E(l,k)>0}} E(l,k) \tag{12}$$

$v(l)$ displays sharp maxima at transients and note onsets, i.e., those instants where the positive energy flux is large. Figure 4 shows an example for a trumpet signal. The top figure is the time–frequency representation, in the middle the SEF $E(l,k)$ and in the bottom the detection function.

### 2.2. Thresholding and peak-picking

The shape of the detection function bears a great importance. In an ideal case, at those time instants where phenomenal accent occur the detection function would display well-localized narrow peaks whose magnitudes are proportional to the intensity change. A simple peak-picking above a fixed threshold would be enough to find onset locations. In practice, the detection function tends to be noisy for a number of reasons. In addition, its dynamics tend to vary considerably over the range of real world signals and in many cases, within short segments of signals there may be a range of different types of onsets. For these reasons, prior to peak-picking, noise removal and dynamic thresholding operations are required. Under the risk of blurring the attacks and losing time resolution, a careful low-pass filtering can be used to remove noise components. For the ease of convenience, we decided to use the above mentioned solution of filtering using the exponential decay function $s(m)$

$$\hat{v}(l) = \sum_m v(m)s(l-m). \tag{13}$$

The dynamic threshold was computed using the method suggested by Bello [2] and previously employed for detecting impulsive noise in audio signals. The basic principle is to find the median of a signal within a sliding analysis window, above which all peaks are selected to pass to the peak-processing stage of the algorithm. Each value of the dynamic threshold $\theta(l)$ is given by

$$\theta(l) = \mathcal{C} \cdot \text{median}(g_m) \tag{14}$$

where $g_m = \{\hat{v}_{m-P}, \ldots, \hat{v}_m, \ldots, \hat{v}_{m+P}\}$ and $\mathcal{C}$ is a predefined scaling factor to artificially rise the threshold curve slightly above the steady state level of the detection function. To ensure accurate detection, the length of the median filter $2*P+1$ must be longer than the average width of the detection function peaks. The bottom part of Figure 4 shows the adaptive threshold calculated for the above mentioned trumpet signal.

A peak-processing stage selects onsets candidate peaks above the adaptive threshold and discards those being too small in a 50 ms range around a larger peak. Finally, peak-picking is carried out.

## 3. PERFORMANCE ANALYSIS

The performance of the presented onset detection system was evaluated using a public database of 670 hand-labeled onsets cross-validated by three different annotators [8]. The signals were selected to comprise a large variety of musical instruments including drums and vocals, a wide dynamic and a wide pitch range. The algorithm parameters, such as: FFT size, differentiator filter order and median filter length, were globally set using a trial-and-error method to search in the space of possible values the ones that maximize a global score. The proposed system was compared to the spectral difference proposed in [4], that is

$$\hat{d}(l) = \sum_k H(|\tilde{X}(l,k)|) - H(|\tilde{X}(l-1,k)|) \tag{15}$$

where $H(x) = \text{arcsinh}(x)$. In addition, $\hat{d}(l)$ is the half-wave rectified version of the signal, $d(l) = \text{HWR}(\hat{d}(l))$. The procedure

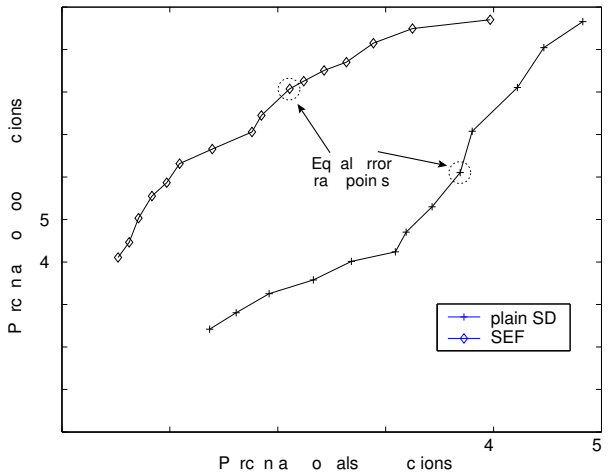


**Fig. 5**. ROC curves obtained by variating the adaptive threshold scaling factor. Percentage of good detections *vs*. percentage of false detections for the SEF and the plain spectral difference.

described in section 2.2 of smoothing, adaptive thresholding and peak processing was also applied to this algorithm.

The performance analysis for both methods are presented in the form of ROC (Receiver Operating Characteristic) curves in Figure 5. Results are shown as percentages of good detections *vs.* false detections for each algorithm. For the analysis, onset candidates found by any of the algorithms were marked as correct if the difference between the candidate and a real onset is shorter than 50 ms, otherwise they were marked incorrect.

Although the SEF enhancement largely outperformed the plain spectral difference approach, the proposed system is still far from being optimal. In general, the detection function $v(l)$ estimated by the algorithm was considered very effective. The main drawback of the system lies on the adaptive thresholding method employed. In fact, the same approach is used for all types of music signals. It may be more appropriate to develop signal specific approaches to better take into account the large variety of music signals. Under the current system, obtaining additional information from the target application (pitch tracking for example) prior to thresholding should make onset detection more robust.

In a more detailed analysis, the algorithm displayed good performance for the wind instruments (flute, sax and trumpet), although the false detection rate was the highest among all categories. It's interesting to notice that, in many cases, these false detections were related to the finger movement pressing the instrument keys, sometimes inaudible but still detectable. For the bowed-string instruments (violin, cello), soft attacks are hard to detect, since the appearance of new frequency components is very gradual resulting in widespread and small peaks. For the percussive music, most of the false detections were due to artifacts in the detection function caused by vocals. The system also proved to lose efficiency when processing signals with large reverberation levels.

The global complexity of the algorithm per each analysis frame is $O(N\log_2(N))$ for an FFT of size $2N$ plus $3N+N\cdot 2L$ multiplications, $3N+N\cdot(2L-1)+N-1$ additions and $N$ divisions for smoothing with the decaying exponential window $s(m)$ and using a differentiator filter $h(m)$ of order $2L$. The storage requirement is of $2L-1$ FFTs. Even if the system is slightly more complex than the plain spectral difference, the proposed algorithm can easily run in real time, demanding access to only a few tenths of milliseconds ahead. The implementation under Matlab© is available for free under request to the first author. In addition, specific examples of the algorithm described can be consulted by internet at the address www.tsi.enst.fr/∼malonso/icme05.

## 4. CONCLUSIONS

In this paper, we have proposed an automatic onset detection system that can be used as a front-end for a large number of computer music applications. The presented algorithm is an enhancement to a previously existing scheme. To find the onsets, a perceptually plausible power envelope is calculated in a band-wise fashion. Then its derivative is computed using an efficient differentiator filter. Contributions from all bands are integrated in frequency to obtain a function that bears onset locations as peaks. The performance displayed by the algorithm is considered satisfactory for a wide range of music signals and attacks. The determination of a better discriminating threshold should considerably improve the algorithm performance. Peak-picking is a high risk operation, direct processing of the detection function combined to target application information should substantially decrease false detections and give place to a more robust system. Real time implementations are possible, requiring only a few milliseconds of future information.

## 5. REFERENCES

[1] F. Lerdahl and R. Jackendoff, *A generative theory of tonal music*. MIT Press, Cambridge, Massachusetts, 1983.

[2] J. P. Bello, "Towards the automated analysis of simple polyphonic music: A knowledge based approach," Ph.D. dissertation, Queen Mary University of London, 2003.

[3] A. P. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE ICASSP*, 1999.

[4] J. Laroche, "Efficient tempo and beat tracking in audio recordings," *Audio Eng. Soc.*, vol. 51, no. 4, 2004.

[5] M. Dvornikov, "Formulae of numerical differentiation," math.NA/0306092, 2003.

[6] R. Meddis, "Simulation of auditory-neural transduction:Further studies," *J. Acoust. Soc. Am.*, vol. 83, no. 3, pp. 1056–1063, March 1988.

[7] B. C. Moore, Ed., *Hearing*, 2nd ed. Academic Press, 1995.

[8] P. Leveau, L. Daudet, and G. Richard, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *Proc. of the ISMIR*, 2004.

## 6. ACKNOWLEDGEMENTS

The authors would like to thank Pierre Leveau and Laurent Daudet for their valuable comments and for making available some of their code for testing purposes.