Note Onset Detection in Musical Signals

Shubhransh Singhvi—Charchit Gupta

1 Abstract

In music analysis, the beginning of events in a music signal (i.e. sound onset detection) is important for tasks such as sound segmentation, beat recognition and automatic music transcription.
The report presents a comparison of six different onset detection techniques

2 Introduction

2.1 Musical Introduction and Motivation

Music is to a great extent a change based phenomenon for both performer and listener. Without change there can be no musical meaning.
Traditionally, written music uses note symbols to indicate the pitch, onset time, and duration of each sound to be played. The loudness and the applied musical instruments are not specified for individual notes but are determined for larger parts. An example of the traditional musical notation is shown below:

Figure 1: An excerpt of a traditional musical notation

Due to the digitization of music signals, onset detection can be automated. The automatic detection of events in audio signals gives new possibilities in a number of music applications including content delivery, compression, indexing and retrieval.

2.2 Definitions

2.2.1 Onset

One can define onset as the start of a musical note (not restricting notes to those having a clearly defined pitch) (Dixon, 2006), or as a single instant chosen to mark the temporally extended transient (Bello et al., 2005). The transient can be understood as a short-time interval in which a significant energy change occurs in the signal (Bello et al.,2005; Klapuri & Davy, 2006). For the ideal case of a single musical note, one can see a clear definition of these concepts in the schema present in figure2.

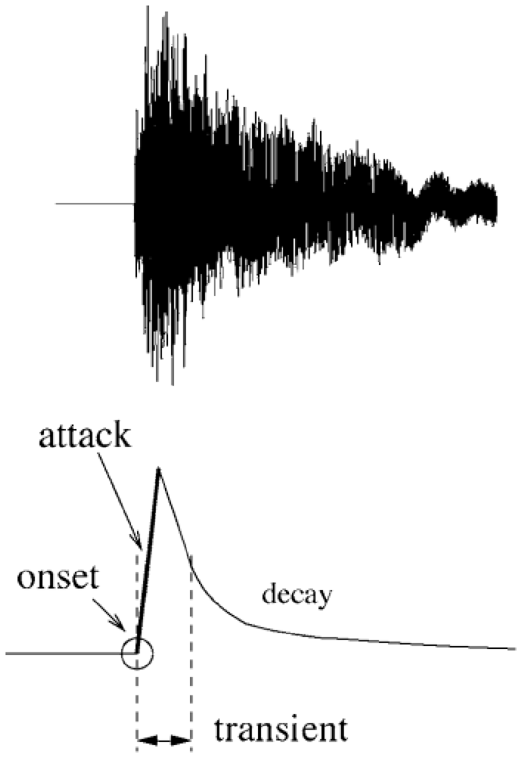

Figure 2: Attack, onset and transient in a single note

2.2.2 Onset Classes

Sounds can broadly be classified into two classes. Harmonic sound on the one hand side is what we perceive as pitched sound and what makes us hear melodies and chords. Percussive sound on the other hand is noise-like and usually stems from instrument onsets like the hit on a drum or from consonants in speech.

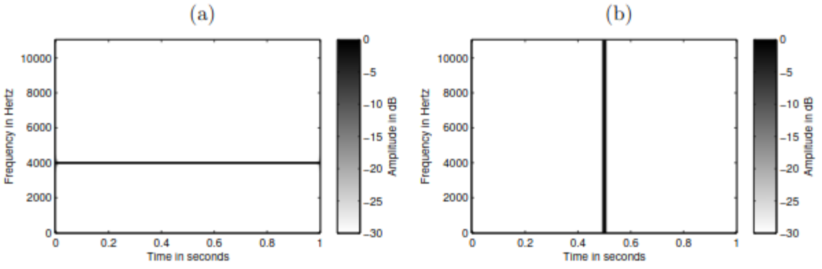

Figure 3: Spectrogram of: (a)an ideal harmonic signal (b)an ideal percussive signal

Hence, we can distinguish the following onset classes:
- Non-Pithced Percussive(NPP)
- Pithced Percussive(PP)
- Pithced Non-Percussive(PNP)
- Complex Mixtures(Mix.)

One can think of Complex Mixture as any polyphonic music where several instruments are played together, something that happens, for instance, in a rock or pop song. The NPP onsets are the ones typically produced by percussion instruments such as drums or cymbals,while the PP onsets are those that have a percussive characteristic but, nonetheless, still maintain a well defined pitch; these onsets appear, for instance, when a piano is playing.Finally, the PNP onsets are those that do not have percussive characteristics and have a very well defined pitch; this category contains onsets from instruments such as bowed strings or wind instruments.

2.3 General Scheme of Onset detection algorithms

It is generally not possible to detect onsets directly without first quantifying the time-varying “transientness” of the signal.
Audio signals are both **additive** and **oscillatory**. Therefore, it is not possible to look for changes simply by differentiating the original signal in the time domain; this has to be done on an intermediate signal that reflects, in a simplified form, the local structure of the original. Such a signal is referred as a **detection function**.
Figure 4 illustrates the procedure employed in the majority of onset detection algorithms: from the original audio signal,
which can be pre-processed to improve the performance of subsequent stages,
a detection function is derived at a lower sampling rate,
to which a peak-picking algorithm is applied to locate the onsets.

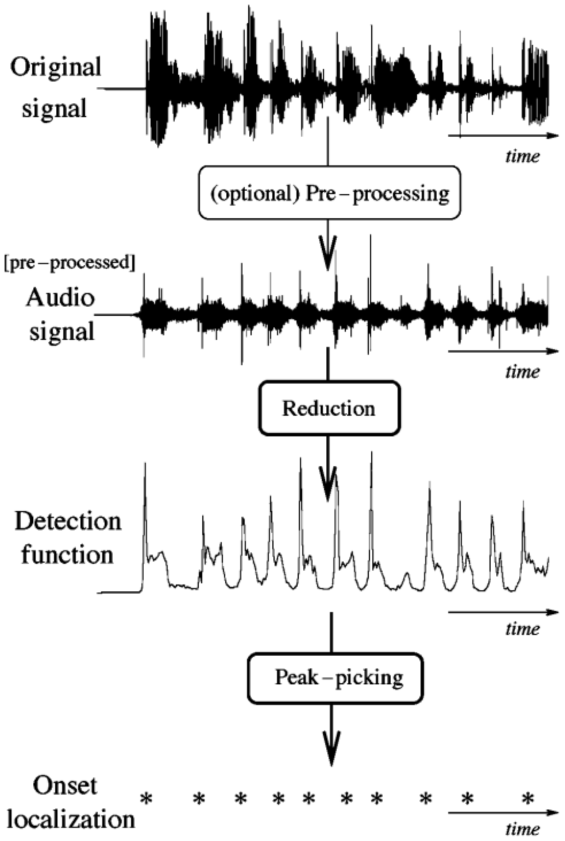

Figure 4: Flowchart of a standard onset detection algorithm

3 Solution Approach and Simulations

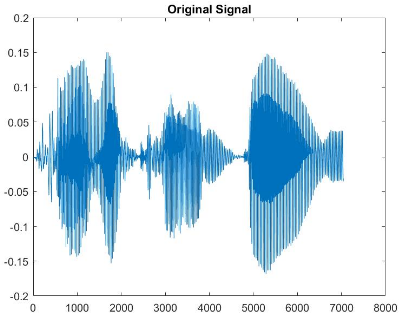

Figure 5: original signal

3.1 STFT

One of the most intuitive and straightforward approaches to perform Time-Frequency analysis and synthesis is to use a pair of localized forms of the Fourier transform termed Short-Time Fourier Transform (STFT) and Inverse Short-Time Fourier Transform (ISTFT).
The STFT of the signal x(n) is given by $X[k,l]$:

$$x_l[m] = x[m + lH]w[m] \tag{1}$$

$$X[k,l] = \frac{1}{M}\Sigma_{m=1}^{K} x_l[m]e^{-j2\pi\frac{mk}{K}} \tag{2}$$

Spectrogram of x(n):

$$s(l,k) = |X[l,k]|^2 \tag{3}$$

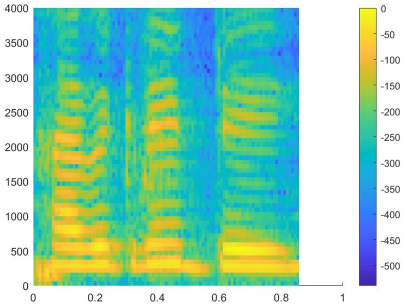

Figure 6: spectrogram

3.2 High Frequency Method

From the STFT, one can define an energy envelope function by summing the power of frequency components in the spectrogram

$$E(l) = \frac{1}{M}\Sigma_{k=-\frac{M}{2}}^{\frac{M}{2}-1}|X[l,k]|^2 = \Sigma_{k=-\frac{M}{2}}^{\frac{M}{2}-1} s(l,k) \tag{4}$$

By observing that an energy increase in one or more frequency bands can be a simple indicator of an onset one can notice that an onset has a more intense energy in the bands in which the interference with other simultaneous components is smaller, a situation which typically occurs in the high-frequencies region. This fact can be exploited by weighting each STFT bin with a factor proportional to its frequency. Hence, by summing all weighted bins, one obtains a function called HFC, that can be used as detection function:

$$HFC(l) = \frac{1}{M}\Sigma_{k=-\frac{M}{2}}^{\frac{M}{2}-1} W_k|X[l,k]|^2 \tag{5}$$

Masri proposed a linear wieghting of $W_k = |k|$
This method works well for percussive onsets, it shows weaknesses for other onset types.

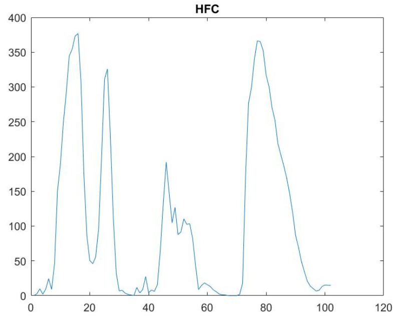

Figure 7: HFC

3.3 Spectral Difference Method

This is a more general approach based spectral changes of the signal, and is related to the formulation of the detection function as a “distance” between successive STFT, treating them as points in an N-dimensional space.

3.3.1 L1-Norm based detection function

If L1-norm is used as a distance function, then, the detection function is called as Spectral Flux.
With the L1-norm it becomes:

$$SF(l) = \Sigma_{k=-M/2}^{M/2-1} H(|X_k(l)| - |X_k(l-1)| \tag{6}$$

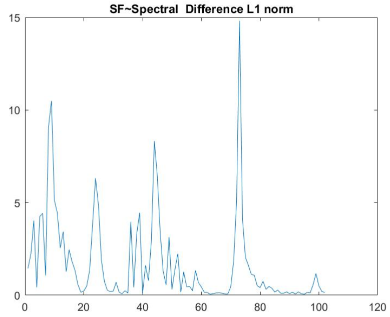

Figure 8: SF

3.3.2 L2-Norm based detection function

If L1-norm is used as a distance function, then, the detection function is called as Spectral Flux. With the L1-norm it becomes:

$$SD(l) = \Sigma_{k=-M/2}^{M/2-1}\{H(|X_k(l)| - |X_k(l-1)|\}^2 \tag{7}$$

where $H(x) = \frac{x+|x|}{2}$ is called the half-wave rectifier function and has the purpose of eliminating negative differences. In this way, it ignores offsets and sticks to onsets.

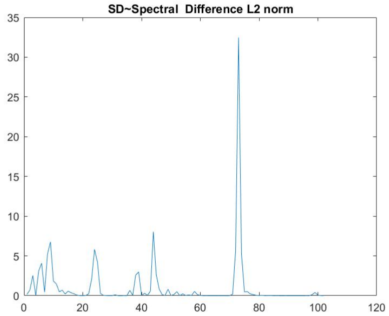

Figure 9: SD

The methods using Spectral Difference or Spectral Flux give very good results in finding NPP onsets.

3.4 Phase Deviation

The methods we have seen so far use the magnitude of the spectrum as their source of information, however, it's possible to make use of the phase spectra. This type of analysis is also important, because much of the temporal structure of a signal is contained in the phase spectrum.

The change of the phase($\phi_k(l)$) in a STFT frequency bin is a rough estimate of its instantaneous frequency, and can be used as indicator of an onset. One can write the instantaneous frequency as:

$$\phi'_k(l) = \phi_k(l) - \phi_k(l-1) \tag{8}$$

And the variation of the instantaneous frequency as:

$$\phi''_k(l) = \phi'_k(l) - \phi'_k(l-1) \tag{9}$$

During a transient region, the instantaneous frequency is not usually well defined and hence $\phi''_k(l)$ will tend to be large.

During the steady-state part of a sound, deviations tend to zero, thus the distribution is strongly peaked around this value. During attack transients,$\phi''_k(l)$ values increase, widening and flattening the distribution.

3.4.1 PD

A simple measure of the spread of the distribution is calculated as

$$PD(l) = \frac{1}{M} \Sigma_{k=-M/2}^{M/2-1} |\phi_k''(l)| \tag{10}$$

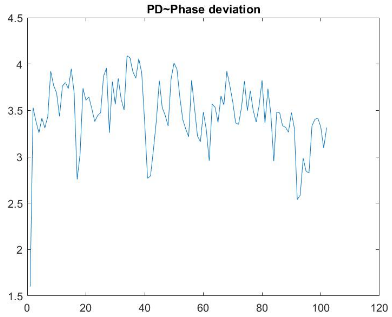

Figure 10: PD

3.4.2 WPD

The PD considers all frequencies equally, so Dixon (2006) proposed weighting the frequency bins by their magnitude, in order to obtain a new onset detection function, WPD:

$$WPD(l) = \frac{1}{M} \Sigma_{k=-M/2}^{M/2-1} |X_k(l) \phi_k''(l)| \tag{11}$$

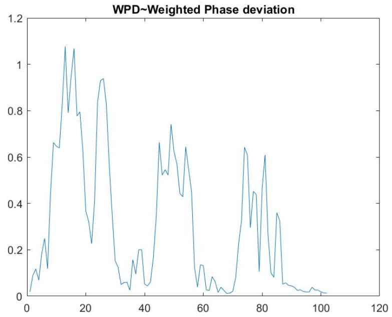

Figure 11: WPD

These methods based on phase deviation tend to give better results on PNP onsets than the methods that use the spectral magnitude, though, in NPP onsets the results are not as good as the ones obtained with the other type of methods.

3.5 Peak-Picking

If the detection function has been suitably designed, then onsets or other abrupt events will give rise to well-localized identifiable features in the detection function. Commonly, these features are local maxima (i.e., peaks), generally subject to some level of variability in size and shape, and masked by 'noise', either due to actual noise in the signal, or other aspects of the signal not specifically to do with onsets, such as vibrato. Therefore a robust peak-picking algorithm is needed to estimate the onset times of events within the analysis signal.

3.5.1 Thresholding

It is common to define a threshold that separates event-related and non-event-related peaks.

- **Constant Threshold**

The first approach is to define a constant threshold,δ, and, in this case, onsets would be peaks where the detection function,d, is bigger than the threshold:

$$d(l) \geq \delta \tag{12}$$

Since music typically exhibits great dynamics, constant thresholds usually give weak results, so it is common to use adaptive thresholds. An adaptive threshold can be constructed in several ways.

- **Threshold Function based on the local mean or local median**

$$\tilde{\delta}(n) = \delta + \lambda * mean(|d(n-M)|, ..., |d(n+M)|) \tag{13}$$

or

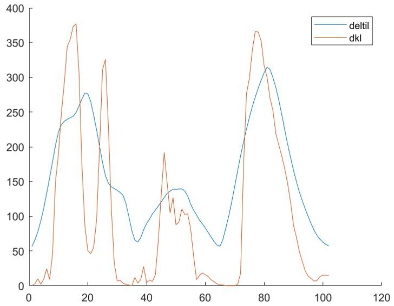

Figure 12: mean thresholding

$$\tilde{\delta}(n) = \delta + \lambda * median(|d(n-M)|, ..., |d(n+M)|) \tag{14}$$

where λ and δ are positive constants, that can be tweaked, and M is the size of the window

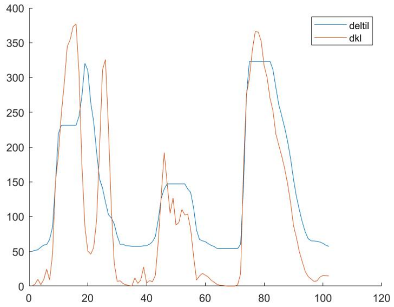

Figure 13: median thresholding

are around each of the points of the detection function.
These threshold functions based on the mean and on the median are the most robust to signal dynamics.
The parameters used in thresholding have a large impact on the final results, mainly in the ratio of false positives to false negatives.

3.5.2 Normal Peak-Picking

Picking the onsets,o(n), is reduced to the identification of local maxima above the defined threshold:

$$o(n) = \begin{cases} 1 & \text{if} \quad d(n) > \tilde{\delta}(n) \quad \text{and} \quad d(n-w) \leq d(n) \leq d(n+w) \\ 0 & otherwise \end{cases} \tag{15}$$

Number of onsets = 4

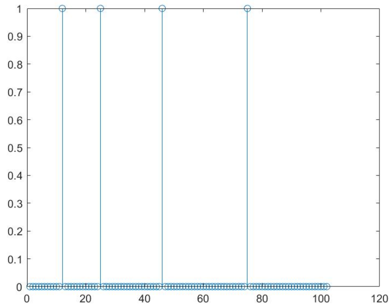

Figure 14: Modified peak-picking

4 References

Original paper:
J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. B. Sandler, ”A tutorial on onset detection in music signals,” in IEEE Transactions on Speech and AudioProcessing, vol. 13, no. 5, pp. 1035-1047, Sept. 2005.
STFT and spectrogram:
Zhivomirov, H.. (2019). On the development of STFT-analysis and ISTFT-synthesis routines and their practical implementation. TEM Journal. 8. 56-64. 10.18421/TEM81-07.
HFC:
P. Masri, “Computer Modeling of Sound for Transformation and Synthesis of Musical Signal,” Ph.D. dissertation, Univ. of Bristol, Bristol, U.K., 1996.
PD:
C. Duxbury, M. Sandler, and M. Davies, “A hybrid approach to musical note onset detection,” in Proc. Digital Audio Effects Conf. (DAFX,’02), Hamburg, Germany, 2002, pp.33–38.
SD, SF, WPD, NWD:
Dixon, S. (2006, September). Onset Detection Revisited. InProceedings of the Int. Conf. onDigital Audio Effects (DAFx-06)(pp. 133–137)
Constant Threshold:
Klapuri, A., Eronen, A., & Astola, J. (2006). Analysis of the meter of acoustic musicalsignals.IEEE Transactions on Audio, Speech, and Language Processing,14(1), 342–355
Mean Threshold:
Kauppinen, I. (2002). Methods for detecting impulsive noise in speech and audio signals. In14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat.No.02TH8628)(Vol. 2, pp. 967–970). IEEE.