COMPLEX DOMAIN ONSET DETECTION FOR MUSICAL SIGNALS

Chris Duxbury, Juan Pablo Bello, Mike Davies and Mark Sandler

Department of Electronic Engineering
Queen Mary, University of London
Mile End Road, London E1 4NS, UK

ABSTRACT

We present a novel method for onset detection in musical signals. It improves over previous energy-based and phase-based approaches by combining both types of information in the complex domain. It generates a detection function that is sharp at the position of onsets and smooth everywhere else. Results on a hand-labelled data-set show that high detection rates can be achieved at very low error rates. The approach is more robust than its predecessors both theoretically and practically.

1. INTRODUCTION

Temporal segmentation of audio into note events is useful for a range of audio analysis, editing and synthesis applications, such as automatic transcription [1], non-linear time-scaling and pitch-shifting as in [2], and content analysis. Figure 1 shows a simple case of two piano onsets, illustrating the energy increase, short duration and instability related to note onset transients, as well as the stability of the steady-state part. The wide range of signals and onset types can be considerably more complex than this example, but these phenomena are common to most.

Almost all onset detection algorithms can be separated into two distinct parts. The first of these, often called the *detection function*, converts the signal from its time domain samples into a function which is more effective in locating onset transients. The second part of any onset detection algorithm is often called the *peak picking* stage, and involves locating points in the detection function which correspond to onset transients.

A strong detection function will typically have sharp peaks located at transients, and few spurious peaks located elsewhere. The more this is the case, the more robust the detection function is to the peak picking algorithm used. For this reason , the majority of this paper is concerned with the detection function generation stage.

The peak picking stage should be effective in selecting only those peaks corresponding to note onsets. Therefore, simply choosing all peaks is only effective in the unlikely case of a perfect detection function. Effective thresholding of the detection function to ignore spurious peaks is a common problem in the peak picking stage. Section 4 will briefly discuss peak picking, however this work is more concerned with detection function generation which is robust to the peak picking method used.

Typically, note onset detection schemes use energy based approaches, often involving frequency weighting [3]. In recent years, this has been extended to include sub-band schemes such as [4, 5]. In [6], a phase based approach to onset detection was proposed. This approach offers clear improvements to those signals which have softer, less percussive onsets. The idea was extended on [7] where a combined phase and energy based approach was proposed. Here, we bind the previous work by developing a complex domain approach to note onset detection.

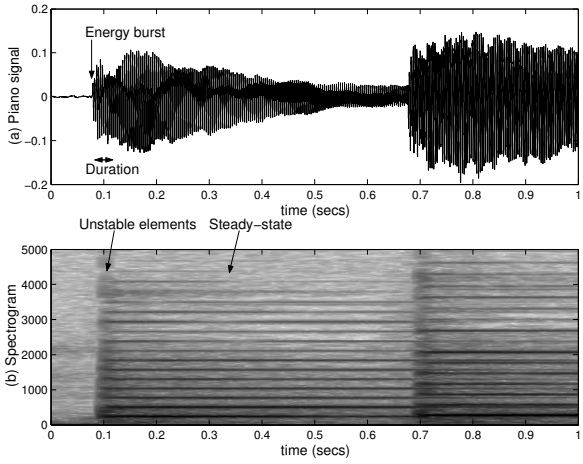

Figure 1: Sequence of two piano notes (a) and the corresponding spectogram (b).

2. PREVIOUS APPROACHES TO NOTE ONSET DETECTION

2.1. Energy-Based Onset Detection

A new note will always lead to an increase in signal energy. In the case of strong percussive note attacks, such as drums, this increase in energy will be very sharp. For this reason, energy has proved to be a useful, straightforward, and efficient metric by which to detect percussive transients, and therefore certain types of note onset. If we consider the L_2 norm squared energy of a frame of the signal, $x(m)$:

$$E(m) = \sum_{n=(m-1)h}^{mh} |x(n)|^2 \qquad (1)$$

where h is the hop size, m the hop number and n is the integration variable. Taking the first derivative of $E(m)$ produces a detection function from which peaks may be picked to find onset locations. This is one of the simplest approaches to note onset detection. This idea can be extended to consider frames of an FFT. Consider a

time-domain signal $s(mh)$, whose STFT is given by:

$$S_k(m) = \sum_{n=-\infty}^{\infty} s(n)w(mh-n)e^{-j2\pi nk/N} \quad (2)$$

where $k = 0, 1, \ldots, N-1$ is the frequency bin index and $w(n)$ is a finite-length sliding window. It follows that the amplitude difference is then:

$$\delta S = \sum_{k=1}^{N} |S_k(m)| - |S_k(m-1)| \quad (3)$$

2.2. Phase-based onset detection

Intuitively, Fourier analysis proposes that a signal can be represented by a group of sinusoidal oscillators with time-varying amplitudes, frequencies and phases. During the steady-state part of the signal these oscillators will tend to have stable amplitudes and frequencies. Therefore, the phase of the k^{th} oscillator at a given time n could be easily predicted according to:

$$\tilde{\varphi}_k(n-1) - \tilde{\varphi}_k(n-2) = \tilde{\varphi}_k(n) - \tilde{\varphi}_k(n-1) \quad (4)$$

where the $\tilde{\varphi}$ operator denotes phase unwrapping. This implies that the actual phase deviation between the target and the real phase values is given by the term:

$$d_\varphi = \text{princarg}[\tilde{\varphi}_k(n) - 2\tilde{\varphi}_k(n-1) + \tilde{\varphi}_k(n-2)] \quad (5)$$

where princarg maps to the $[-\pi, \pi]$ range. d_φ will tend to zero if the phase value is accurately predicted and will deviate from zero otherwise. The latter is the case for most oscillators during attack transients.

This can be extrapolated to the distribution of deviations for all oscillators in one analysis frame. During the steady-state part of a signal most values will be concentrated around zero creating a sharp distribution. On the other hand, during attack transients the distribution will be wide and less sharp. By measuring the spread of the distribution an accurate onset detection function can be constructed [6].

2.3. Combining phase and energy based approaches

While energy-based methods are straightforward, and thus widely used, they rely on the presence of pronounced energy increases for all events in music. However this is not always the case, especially with complex mixtures when overlapping between notes is common. Phase-based approaches offer an alternative to this, increasing effectiveness for less salient onsets. However, these methods are susceptible to phase distortion and to the variations introduced by the phase of noisy components.

In [7] a method was proposed that combined both the energy and the phase approaches. It made use of the similar behavior of the distributions of phase deviations and of spectral magnitude differences. Measures of spread per frame for each distribution were obtained as:

$$\eta(n) = \text{mean}(f_n(|x|)) \quad (6)$$

where $f(x)$ is the probability density function of our data set. Then they were multiplied, emphasising the phase characteristic of those components most relevant for the analysis. The method compensated for instabilities in either approach, and produced sharper peaks for detected onsets. Results consistently outperformed both the energy and the phase-based methods.

3. COMPLEX DOMAIN ONSET DETECTION

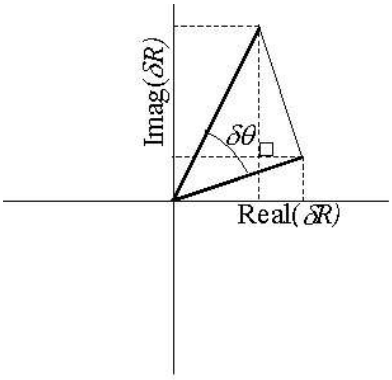

Figure 2: Phasor diagram in the complex domain showing the phase deviation between target and current vector, and the Euclidean distance between them.

By definition, for locally steady state regions in audio signals, it holds that the frequency and amplitude remain constant. In the previous sections it has been shown that by inspecting changes in either frequency and amplitude onset transients can be located. However, we can simultaneously consider the effect of both variables by predicting values in the complex domain. It can be assumed that, in its polar form, the target value for an FFT bin is given by:

$$\hat{S}_k(m) = \hat{R}_k(m)e^{j\hat{\phi}_k(m)} \quad (7)$$

where the target amplitude $\hat{R}_k(m)$ corresponds to the magnitude of the previous frame $|S_k(m-1)|$, and the target phase $\hat{\phi}_k(m)$ can be calculated as the sum of the previous phase and the phase difference between preceding frames:

$$\hat{\phi}_k(m) = \text{princarg}[2\tilde{\varphi}_k(m-1) - \tilde{\varphi}_k(m-2)] \quad (8)$$

We may then consider the measured value in the complex domain from the FFT:

$$S_k(m) = R_k(m)e^{\phi_k(m)} \quad (9)$$

where R_k and ϕ_k are the magnitude and phase for the kth bin of the current STFT frame. By measuring the Euclidean distance between target and current vectors in the complex space, as shown in Figure 2, we can then quantify the stationarity for the k^{th} bin as:

$$\begin{aligned}\Gamma_k(m) \quad = \quad & \{[\Re(\hat{S}_k(m)) - \Re(S_k(m))]^2 + \cdots \\ & [\Im(\hat{S}_k(m)) - \Im(S_k(m))]^2\}^{\frac{1}{2}}\end{aligned} \quad (10)$$

Summing these stationarity measures across all k, we can construct a frame-by-frame detection function as:

$$\eta(m) = \sum_{k=1}^{K} \Gamma_k(m) \quad (11)$$

Equation 11 can be simplified by mapping $\hat{S}_k(m)$ onto the real axis (forcing $\hat{\phi}_k(m) = 0$), such that:

$$\hat{S}_k(m) = \hat{R}_k(m) = R_k(m-1) \tag{12}$$

This implies rotating the phasors in Fig. 2, so that $S_k(m)$ can be represented using the phase deviation (Eq. 5):

$$S_k(m) = R_k(m)e^{jd_{\varphi k}(m)} \tag{13}$$

We now consider the difference between this complex domain prediction approach, and the basic amplitude difference measure for the kth bin:

$$\delta S_k(m) = \hat{R}_k(m) - R_k(m) \tag{14}$$

With the mapping onto the real axis of $\hat{S}_k(m)$, equation 10 becomes:

$$\Gamma_k(m) = \{[\hat{R}_k(m) - \Re(S_k(m))]^2 + \Im(S_k(m))^2\}^{\frac{1}{2}} \tag{15}$$

which becomes:

$$\begin{aligned}\Gamma_k(m) &= \{[\hat{R}_k(m) - R_k(m)cos\big(d_{\varphi k}(m)\big)]^2 + \cdots \\ &\quad [R_k(m)sin\big(d_{\varphi k}(m)\big)]^2\}^{\frac{1}{2}}\end{aligned} \tag{16}$$

This can then be expanded:

$$\begin{aligned}\Gamma_k(m) &= \{\hat{R}_k^2(m) - 2\hat{R}_k(m)R_k(m)cos(d_{\varphi k}(m)) + \cdots \\ &\quad R_k^2(m)sin^2(d_{\varphi k}(m)) + \cdots \\ &\quad R_k^2(m)cos^2(d_{\varphi k}(m))\}^{\frac{1}{2}}\end{aligned} \tag{17}$$

Simplifying, we obtain:

$$\begin{aligned}\Gamma_k(m) &= \{\hat{R}_k(m)^2 + R_k(m)^2 - \cdots \\ &\quad 2\hat{R}_k(m)R_k(m)cos(d_{\varphi k}(m))\}^{\frac{1}{2}}\end{aligned} \tag{18}$$

For the case of $d_{\varphi k}(m) = 0$:

$$\begin{aligned}\Gamma_k &= \{\hat{R}_k^2(m) + R_k^2(m) - 2\hat{R}_k R_k\}^{\frac{1}{2}} \\ &= \hat{R}_k(m) - R_k(m)\end{aligned} \tag{19}$$

Therefore $\Gamma_k(m)$ is only equal to $\delta S_k(m)$ where $d_{\varphi k}(m)$ is equal to zero, or when the phase prediction is "good". In that case, only the energy difference is being taken into account. In the case of $d_{\varphi k}(m) \neq 0$, there is the additional term taking the phase deviation from the prediction into account.

$\eta(m)$ constitutes an adequate detection function showing sharp peaks at points of poor stationarity. Figure 3 depicts the detection function for a section of a guitar signal. The figure also gives examples of phase and amplitude used individually. The complex domain approach is clearly less noisy, therefore simplifying the task of peak-picking and allowing a more robust detection.

4. PEAK PICKING

The thresholding of onset detection functions is problematical for a number of reasons. Firstly, the detection functions tend to be noisy, unless they are extensively low pass filtered, leading to a poorer time resolution, and loss of weaker transients. Secondly, detection function magnitudes tend to vary considerably over the range of real world signals. Further to this, within one short segment of a signal there may be a range of different types of onsets. For these reasons, detection thresholds tend to be set manually in many onset

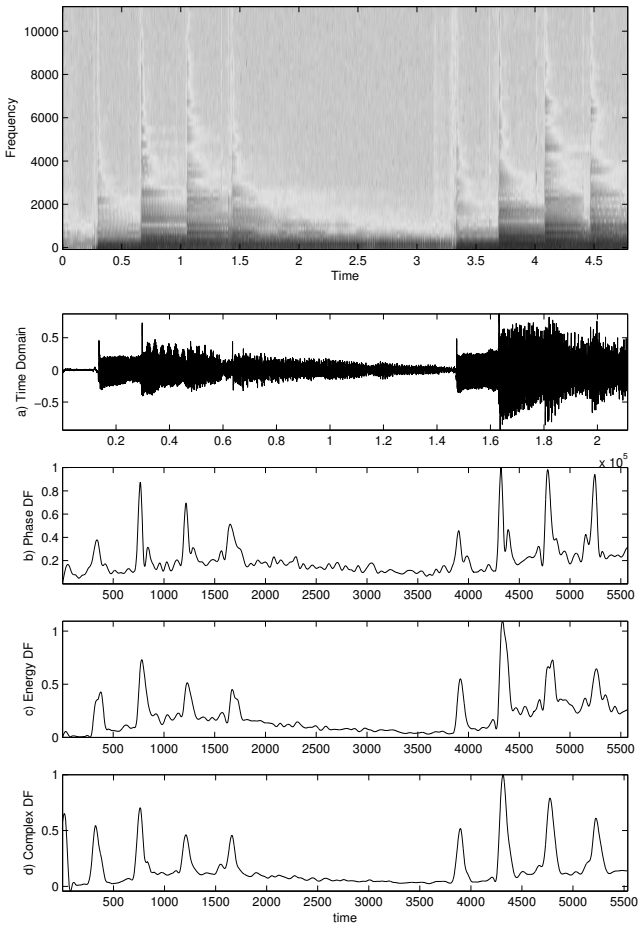

Figure 3: Spectrogram of a music signal and its corresponding time domain representation (a), Phase Based Detection Function (b), Energy Distribution Detection Function (c) and Complex Domain Prediction Detection Function (d).

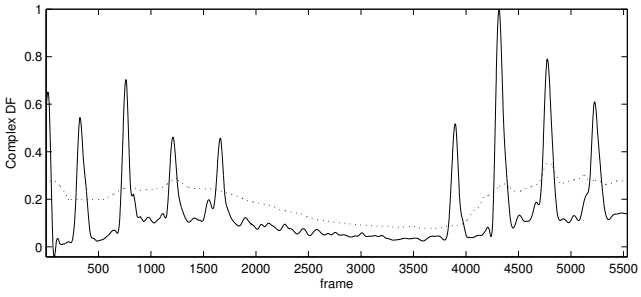

Figure 4: Moving Median Thresholding of Complex Domain Prediction Detection Function.

detection applications. However, there are many cases where this is not practical. For example, when implementing audio effects requiring detection of note onsets, the user should not be required to set an onset detection threshold for each signal. This is also a considerable problem where real time applications are desired.

In peak picking algorithms, thresholding may be set either globally or locally. Whilst global thresholds are a computationally efficient method, the large changes in signal dynamics and content that can occur within the same signal suggest that local thresholding is essential for effective onset detection.

A simple peak-picking algorithm, using a weighted moving average, is used to determine the precise location of note onsets from the detection function. This is based on the thresholding algorithm presented in [8] where it is used for detection of impulsive noise. The basic principle of this is to find the median average of a signal within a sliding analysis window, above which all peaks are selected as onsets.

Each value of the dynamic threshold δ_t, for a H-length sliding analysis is given:

$$\delta_t(m) = C_t \,\text{median}\, \gamma_2(k_m), k_m \in [m - \frac{H}{2}, m + \frac{H}{2}] \quad (20)$$

where C_t is a scaling factor. Figure 4 illustrates the dynamic threshold for the signal shown in figure 3.

5. RESULTS

Experiments were performed on a database of a wide range of polyphonic music examples containing 400 hand-labeled onsets. Figure 5 displays the percentage of false negatives versus the percentage of good detections for different offset values. The ideal curve will rest over the y-axis and the 100% good detection line. It can be seen from the generated curves that the complex domain curve is considerably more robust to the peak picking threshold used than the curves representing phase and energy only distribution detection functions. It also performs consistently better over the entire range.

At the optimum position of the complex domain detection curve, the algorithm achieves an average of 95% good detections, for 2% false negatives. Considering the range and complexity of the musical signals used in this test, this is a remarkably good result.

6. CONCLUSIONS

In general, energy-based onset detection schemes have performed well for audio signals with significant percussive content, or "hard" onsets. Conversely, phase-based onset detection approaches provide a good solution to onset detection for "softer" signals, such as bowed strings. In the complex domain, both phase and amplitude information work together, offering a generally more robust onset detection scheme. This algorithm is both straightforward to implement, and computationally cheap. Despite this, it proves effective for a large range of audio signals.

As this complex domain approach currently performs better on the lower frequency components of the spectrum, it may be beneficial to incorporate it within a mutltiresolution scheme. This has the advantage that high frequency noise bursts may be used to improve time localisation of hard onsets. Since the analysis must be complex in this case, a wavelet based approach would be unsuitable. However, multiresolution Fourier analysis or complex wavelets such as those discussed in [9] may be useful for this purpose.

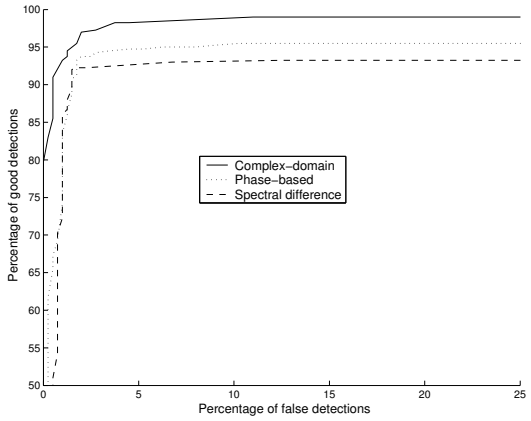

Figure 5: Percentage of good detections versus percentage of false negatives for different weight values and using the complex method

7. REFERENCES

[1] J.P. Bello, *"Towards the Automated Analysis of SImple Polyphonic Music: A Knowledge-based Approach."*, PhD Diss., Queen Mary, University of London, 2003.

[2] C. Duxbury, M. Davies, and M. Sandler, "Improved Time-Scaling of Musical Audio Using Phase Locking at Transients," in *Proc. AES 112th Convention*, 2002.

[3] P. Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals*, PhD Thesis, University of Bristol, 1996.

[4] Xavier Rodet and Florent Jaillet, "Detection and modeling of fast attack transients," in *Proceedings of the International Computer Music Conference*, 2001.

[5] A. Klapuri, "Sound Onset Detection by Applying Psychoacoustic Knowledge," in *Proc. IEEE Conf. Acoustics, Speech and Signal Proceesing (ICASSP'99)*, 1999.

[6] Juan P. Bello and Mark Sandler, "Phase-based note onset detection for music signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP-03*, 2003.

[7] C. Duxbury, J.P. Bello, M. Davies, and M. Sandler, "A combined phase and amplitude based approach to onset detection for audio segmentation," in *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03), London, UK.*, 2003.

[8] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio aignals," in *Proc. DSP2002*.

[9] N. G. Kingsbury, "The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters.," in *Proc. IEEE Digital Signal Processing Workshop, DSP 98, Bryce Canyon UT*, 1998.