PHASE-BASED NOTE ONSET DETECTION FOR MUSIC SIGNALS

Juan Pablo Bello and Mark Sandler

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

ABSTRACT

Note onsets mark the beginning of attack transients, short areas of a note containing rapid changes of the signal spectral content. Detecting onsets is not trivial, especially when analysing complex mixtures. Applications for note onset detection systems include time stretching, audio coding and synthesis. An alternative to standard energy-based onset detection is proposed by using phase information. It is suggested that by observing the frame-by-frame distribution of differential angles, the precise moment when onsets occur can be detected with accuracy. Statistical measures are used to build the detection function. The system is tested and tuned on a database of complex recordings.

1. INTRODUCTION

Providing an adequate explanation of what onset means is a difficult task. Let us start with a simple definition: the onset is the precise moment when a new event begins. This immediately raises the need for a definition of *event* on this context. [1] defines an event as an auditory phenomenon that shows continuity for, at least, the smallest duration that can be perceived. Under this definition, musical events may include expressive features, timbre changes and notes. In this paper we are only concerned with the detection of onsets related to the latter. Applications for note onset detection systems include time stretching, audio coding and synthesis.

A note can be divided into two main components: the steady state and a transients (plus noise) component. We will assume transients and steady state as two separate, sequentially occurring components of an event. Transients precede the steady state of the signal, when the signal is stationary and predictable. They are usually short and contain rapid changes of the signal spectra. Because of the unpredictability of such changes, they are difficult to model. Onsets mark the beginning of notes, thus marking the beginning of attack transient regions.

The task of detecting onsets is not trivial. The boundaries between notes and different types of events are often ill-defined. The physics of the instruments and recording environments create artifacts that can be easily mistaken for onsets. Unsurprisingly, when dealing with polyphonic mixtures, the detection of onsets becomes more complex.

Robust onset detection depends on the understanding of the characteristics of transients. Attack transients are characterised by an steep increase in the note's energy profile. The more impulsive the signal's components the more sudden the increase becomes. Their duration is short, introducing significant changes into the signal. The proliferation of elements whose values are completely unexpected is more likely during the attack. Finally, transients are followed by the steady-state region of the note. Chaotic components followed by stable ones hint at the possibility of a note.

2. PHASE-BASED ONSET DETECTION

Energy-based algorithms are usually fast and easy to implement [2, 3]. However, their effectiveness decreases when transients of the signal are not pronounced (i.e. non percussive sounds) and when energy bursts of different events overlap in polyphonic mixtures. We propose an alternative to this by using phase information. The phase carries all the timing information of an audio signal. It is usually ignored when performing spectral analysis. Transients are well-localised events in time, hence we suggest that phase analysis can return more meaningful results for the detection of new events than solely relying on energy values. Furthermore, as analysing phase implies a type of tonal analysis, changes that are not as noticeable as energy bursts may still be successfully detected as pitch *bursts*.

The proposed onset detection algorithm builds upon a method for phase-based transient / steady-state (TSS) separation [4]. The data produced by the separation algorithm is analysed using statistical methods. By relying on the statistics of our data distribution we intend to generalise our analysis to a large variety of signals. The statistical analysis produces a detection function from which the final results are obtained. In the following, a detailed explanation of the process is provided.

0-7803-7663-3/03/$17.00 ©2003 IEEE

2.1. TSS separation

Let us define a time-domain signal $s(n)$, whose STFT is defined as:

$$S(n,k) = \sum_{g=-\infty}^{\infty} s(g)w(n-g)e^{-j2\pi gk/N} \tag{1}$$

where $k = 0, 1, \dots, N-1$ is the frequency bin index, g is the summation index and $w(n)$ is a finite-length sliding window. $S(n,k)$ can also be defined in terms of its magnitude $|S(n,k)|$ and phase $\varphi(n,k)$ (whose unwrapped version is denoted as $\tilde{\varphi}(n,k)$). Let $n = mR$ where m is the hop number and R is the hop size. According to phase vocoder theory, if presented with a perfect sinusoid in ideal conditions, then we might expect the current phase of the k^{th} bin to be equal to the target phase:

$$\tilde{\varphi}_t(m,k) = \tilde{\varphi}(m-1,k) + \Omega_k R \tag{2}$$

where Ω_k is the frequency of the k^{th} sinusoid. However, real sounds in real conditions fail to comply with this, and instead we might expect our unwrapped estimated phase to be the target phase plus a phase deviation $\tilde{\varphi}_d(m,k)$. This deviation can be calculated as:

$$\tilde{\varphi}_d(m,k) = \mathrm{princarg}[\tilde{\varphi}(m,k) - \tilde{\varphi}_t(m,k)] \tag{3}$$

where $\mathrm{princarg}$ is the *principal argument* function mapping the phase to the $[-\pi, \pi]$ range.

The instantaneous frequency of the k^{th} sinusoid can be defined as the rate of angular rotation, i.e. the unwrapped phase difference $\Delta\varphi(m,k)$ divided by the time between successive frames [5]:

$$f_i(m,k) = \frac{\Delta\varphi(m,k)}{2\pi R} f_s \tag{4}$$

where f_s is the sampling frequency. The unwrapped phase difference is simply the difference between consecutive estimated unwrapped phases.

Consider the behaviour of the individual k^{th} sinusoid. If the sinusoid is stable, it is expected that the instantaneous frequency at hop m will be close to the instantaneous frequency at hop $(m-1)$. Inversely, if the sinusoid is not stable (i.e. when a new, unpredictable event occurs), the variation between these two frequencies increases. This is illustrated in Fig. 1. Note how the instantaneous frequencies of fundamental and the first two partials vary around onset time. This can be expressed by defining an instantaneous frequency difference between consecutive frames $\Delta f_i(m,k)$. the closeness of this difference to zero is an indicator of the stability of the sinusoid. It is proportional to $d\tilde{\varphi}$, geometrically seen as the differential angle between target and current phase. Then, by using equations 2, 3 and 4, we can measure this "differential angle" as:

$$d\tilde{\varphi} = princarg[\tilde{\varphi}(m,k) - 2\tilde{\varphi}(m-1,k) + \tilde{\varphi}(m-2,k)] \tag{5}$$

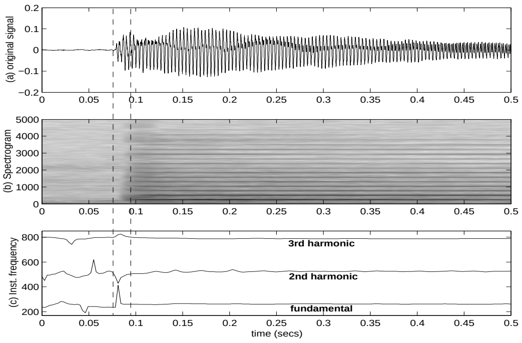


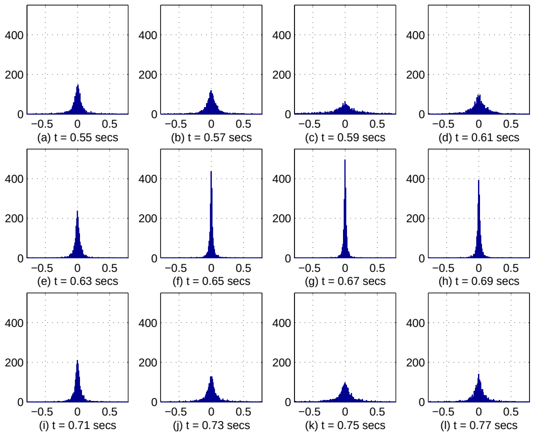
Fig. 1. Instantaneous frequencies (c) of fundamental and first two partials of the note depicted in (a).

By thresholding this angle, effective TSS separation can be achieved [4, 6]. Instead, we suggest that by observing the frame-by-frame distribution of differential angles for all k, a detection function can be created that indicates the precise moment when an onset occurs. Statistical analysis methods are used for this task.

2.2. Statistical analysis

Let us consider all the bin angular changes in one frame as a data set X, such that $x \in X$ are within the $[-\pi, \pi]$ range. The probability density function (PDF) $f(x)$ can be observed by generating a histogram from X. Fig. 2 shows a sequence of PDF's around a note onset. It can be observed (Fig. 2(a) to (b) and (j) to (l)) that in the absence of transients, $f(x)$ closely resembles a normal distribution: unimodal, bell-shaped and symmetrical about the mean. However, when transients occur, due to the non-stationarity of the signal, the difference between target and actual angular position increases, thus the data-set becomes dispersed across the phase range. The spread causes a slight flatness at the top of the distribution, and a decrease of the height of its lobe. This is illustrated in Figure 2(c) and (d), and is particularly noticeable in the former. Immediately after, at the beginning of the steady-state, target and current angular position become closer. The distribution presents a large concentration of zero-phase values, increasing the sharpness and height of $f(x)$. Figures 2(e) to (i), depict this behaviour.

To quantify this observation, the spread of the distributions is measured. The standard deviation σ is used. A sequence of σ values for a music signal is shown in Figure 3(b). It characterises onsets as sharp-peak / deep-valley pairs. However, as can be observed, the standard deviation presents a noisy profile that affects the accuracy of the detection. Alternatively the Interquartile range (IQR) is calculated. For this, the data-set is divided in two equal-sized



(a) t = 0.55 secs (b) t = 0.57 secs (c) t = 0.59 secs (d) t = 0.61 secs (e) t = 0.63 secs (f) t = 0.65 secs (g) t = 0.67 secs (h) t = 0.69 secs (i) t = 0.71 secs (j) t = 0.73 secs (k) t = 0.75 secs (l) t = 0.77 secs

Fig. 2. Sequence of PDF's around onset time

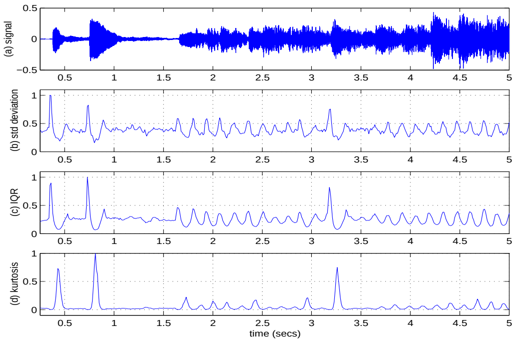


Fig. 3. standard deviation (b), IQR (c) and kurtosis (d) of a test signal (a).

groups at the median. The medians of both, low and high groups are calculated and denoted Q_1 and Q_3 respectively. The IQR is obtained as:

$$IQR \equiv Q_3 - Q_1 \tag{6}$$

Figure 3(c) shows the IQR sequence of the test signal. The profile is less spiky than its predecessor's. The IQR is less sensitive to small variations in the distribution's spread than the standard deviation. These observations are consistent in several test audio files, hence the use of IQR is favoured.

2.3. Shape and kurtosis

In real recordings, phase misalignment between composing sinusoids of a note cause the differential angular distribution to spread. This is critical when notes evolve for a long time, and the partial's incoherence becomes evident. In this case, measuring spread is not robust for onset detection. To overcome this, the analysis of the distribution's spread is complemented with the analysis of the shape of the distribution. The aim is to identify the steady-state that follows the transient.

Kurtosis is the normalised fourth central moment of a distribution and is denoted as γ_2. It measures the flatness or peakedness of the distribution in relation to a normal distribution. The Fisher kurtosis (a common implementation), is defined as:

$$\gamma_2 \equiv \frac{\mu_4(\mu)}{(\mu_2(\mu))^2} - 3 = \frac{\mu_4(\mu)}{\sigma^4} - 3 \tag{7}$$

$\mu_4(\mu)$ denotes the fourth central moment. According to this definition, kurtosis decreases for flat PDFs (platy-kurtic) and increases for sharply peaked PDFs (lepto-kurtic). This is well adjusted to our needs. At the beginning of the steady-state of a note the difference between target and actual phases becomes minimal. Thus, the population concentrates close to the centre of the distribution (increasing the sharpness of its lobe).

The kurtosis successfully represents this characteristic as can be seen in Figure 3(d). We propose that by detecting peaks in the kurtosis profile, the beginning of the steady-state of a note can be accurately detected. Then, by matching each detection to the closest preceding peak in the IQR profile, precise onset times can be pin-pointed.

3. PEAK-PICKING

A peak-picking algorithm is implemented that selects peaks above a dynamic threshold. Each value of the dynamic threshold δ_t is calculated as the weighted median of an H-length section of the kurtosis around the corresponding frame, such that:

$$\delta_t(m) = C_t \,\text{median}\, \gamma_2(k_m), k_m \in [m - \frac{H}{2}, m + \frac{H}{2}] \tag{8}$$

C_t is a predefined weighting value. Low values of C_t increase the number of detections (false included), while high values of C_t make the system more strict. Detected peaks must be separated by, at least, a minimum distance. When several peaks are detected within the minimum distance, only the highest is kept as a valid onset.

4. RESULTS

A small database of commercial recordings was used to test the system's performance. It consists of a number of short segments of different styles of music whose onsets have beeing hand-labelled (a broad style classification can be seen in Table 1). The database contains 324 onsets, both in solo performances and in complex mixtures. Correct matches imply that target and detected onsets are within 50ms of each other (this accounts for the uncertainty, as to precise location, in the hand-labelling process).

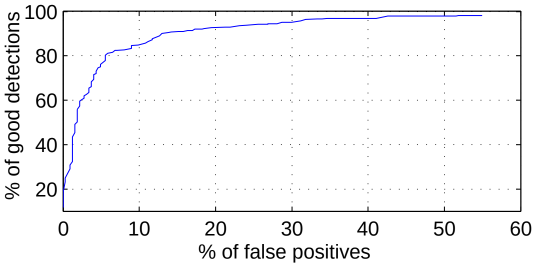


Fig. 4. Percentage of good detections against percentage of false positives for various C_t values

Our first test aims to find the optimal value for C_t and to evaluate the capabilities of our detection function. Fig. 4 shows the relationship between good detections and false positives for different values of C_t. Detections are made over the complete database. It can be observed that good detection rates between 80-90% can be obtained at a cost of around 10% rate of false positives. This is very high when considering the variety and complexity of the signals involved. A system tuned for an specific type of music might generate even better results. By using the value of C_t that corresponds to the "elbow" of the curve, we maximise correct detections.

Having selected an optimal value for C_t, tests are performed over the database. Table 1 shows results according to style. Numbers in the table correspond to percentages of good detection (GD), false positives (FP) and false negatives (FN). Onsets are evenly distributed between different styles. It can be seen that detection rates are high for solo instruments (even for non-percussive instruments such as the violin), with a relatively low cost in false detections. However, as the complexity of the signals increase (i.e. jazz and pop music with vocals), high detection rates are accompanied with higher rates of false positives. A particular case, that of pop rock music with vocals, is singled out as the worst scenario. The features of the singed voice create a number of situations when onsets cannot be properly accounted for (i.e. sibilance, the percussive sound of a 't' at the end of a word, etc) due to phase distortion. The system sometimes fails to resolve these situations. This is a difficulty also present during the hand labelling of those signals.

STYLE	% GD	% FP	% FN
solo violin	81.72	3.80	18.28
solo piano	91.83	6.25	8.16
solo tabla	92.68	7.32	7.32
jazz trio	84.85	9.68	15.15
pop rock	87.50	16.95	12.50
pop rock + vocals	90.32	26.32	9.68

Table 1. Onset Detection Results

5. CONCLUSIONS

A new approach is proposed for the detection of note onsets in music signals. It is an alternative to standard energy-based methods in that it uses the phase of the signal for the detection. Phase vocoder theory is used to generate frame-by-frame statistical distributions of differential angles. The kurtosis and the IQR of the distributions are measured to quantify the distribution's behaviour around onset times. The obtained detection functions allow high detection rates as tested with a database of complex real recordings including both percussive and non-percussive (i.e. violin, voice) instruments. Results could be improved by tuning the system for specific styles of music.

6. ACKNOWLEDGMENTS

Juan Pablo Bello is supported by the OMRAS project, jointly funded by JISC (UK) and NSF (USA). The authors will like to thank Chris Duxbury for his valuable contribution into the ideas here presented and Giuliano Monti for making available some of his code.

7. REFERENCES

[1] D. K. Mellinger, *Event Formation and Separation in Musical Sound*, Ph.D. thesis, Center for Computer Research in Music and Acoustics, Stanford University, 1991, Also Dept of Music Report STAN-M-77.

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

[3] P. Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signal*, Ph.D. thesis, University of Bristol, 1996.

[4] J. Settel and C. Lippe, "Real-time musical applications using the fft-based resynthesis," in *Proceedings of the International Computer Music Conference (ICMC94)*, 1994.

[5] Mark Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14, 1986.

[6] Chris Duxbury, Mike Davies, and Mark Sandler, "Separation of transient information in musical audio using multiresolution analysis techniques," in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland*, December 2001.