Acoustical Onset Detection Using Phase Information

Wen Xue

HCI Lab,
Samsung Advanced Institute of Technology
x.wen@samsung.com

Abstract

Amplitude and phase equally share the information in an audio waveform. While the amplitudes describe the frequency contents of a signal, the temporal structure, which is another crucial property of the waveform, is mostly encoded in the phase angles. In this article we propose a method that makes use of phase spectra for the purpose of onset detection. The way of an onset or offset being encoded in the phase spectra is discussed. Then our method is described in detail, which interprets the existence of onsets into a well-defined pattern that is relatively easy to detect in a sliding window procedure. Phase information is independent on sound level, thus is effective in detecting onsets of very diverse intensities within a short duration. Besides, the phase pattern is generated through long-term integrals, which offers special robustness against noises and fast modulations that have troubled amplitude-based onset detectors. Examples are given on environmental sounds as well as on music to illustrate the effectiveness of the method.

1. Introduction

Through Fourier analysis a signal is represented by its amplitude spectrum and phase spectrum, each sharing half the information. While the amplitude encodes the frequency contents of the signal, the phase stresses on the temporal (spatial) structure. Although important, the phase has been vastly ignored in computational analysis for acoustical signals. Previous work on the retrieval of acoustical information from the phase had focused on frequency-related properties[1-2]. On the contrary, we discuss the temporal information encoded in the phase. This article deals with the task of one-by-one onset detection in sounds. Here one-by-one means to detect every existing onset regardless of macroscopic temporal structures. Previously this type of onset detection is mostly performed by subband methods using amplitude tracks [3-5]. We'll show that the phase spectra can be useful in this task. The proposed method is simple, straightforward and effective, in spite that a thorough understanding of the phase could be complex.

2. Basic Facts

Our work is based on FFT. We start with the simplest case: an impulse signal within the analysis window. This signal is described in time domain by its intensity and offset, or in frequency domain by a flat amplitude spectrum and a linear phase spectrum. Note that the temporal location of the impulse is fully represented by the slope of the linear phase. By subtracting adjacent phase angles, we locate the impulse within the window. This procedure is almost the dual counterpart to that described in [2] which is for frequency estimation. A phase unwrapping procedure is necessary when the calculated phase angles of adjacent spectral samples differ by more than π.

When noise is present, we view the whole phase spectrum as a distorted linear function of frequency. Instead of locally subtracting phase angles, it's effective to use linear fitting to extract a linear phase component. Again the phase unwrapping operation is necessary to restore the phase spectrum in its unwrapped shape. It's essential to note that phase unwrapping relies on the continuity of the phase spectrum, which is debatable around zeroes of the complex spectrum.

The discussions above could be generalized to signals other than impulse. We discuss only the simplest case: a signal with one and only one distinct maximum, of which the impulse is an example. Our goal is to locate the maximum by studying the phase spectrum. First, suppose a piece of signal $s(t)$ with most of its energy distributed at the beginning of the window. We divide $s(t)$ into 2 parts: $s_1(t)$ that takes the significant part with most of the total energy in a short time, and $s_2(t)$ the rest, so that $s(t) = s_1(t) + s_2(t)$ (Figure 1). Since $s_1(t)$ dominates the energy in $s(t)$, it follows that the phase spectrum of $s_1(t)$ approximates that of $s(t)$ according to the parallelogram law of addition. On the other hand, one may regard $s_1(t)$ as a piece of non-zero signal $s'(t)$ padded with zeroes, and its DFT $S_1(k)$ is a smoothly interpolated version of that of $s'(t)$. Thus we conclude that the absolute value of the phase jumps between adjacent samples of $S_1(k)$ (and therefore of $S(k) = H(k)e^{j\varphi(k)} = DFT[s(t)]$) is very small compared to π. Now if we give $s(t)$ a time shift, a linear phase component will be added to $\varphi(k)$. Since this linear phase grows up to $\pm\pi$ for each increase in k when the time shift is large enough, it will finally dominate the phase spectrum, as in the case of impulse signal. Thus we draw the hypothesis, that the linear component of an

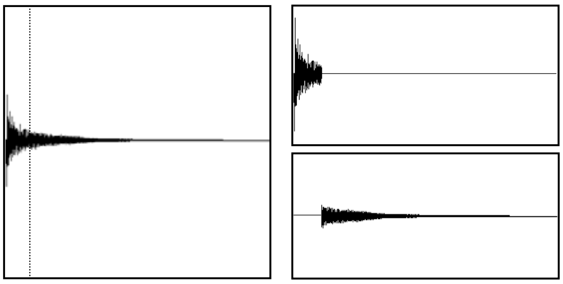

Figure 1. Decomposing a signal according to energy density

unwrapped phase spectrum linearly locates the position of the global maximum of a signal, if there exists one and only one distinct maximum.

Now move to the problem of onset detection. We use a window of fixed length sliding along the time axis for Fourier analysis. Phase spectrum is calculated and unwrapped on each slide. The onset appears first at the end of the window and forms a maximum there. As the window slides on, the maximum moves toward the beginning. As long as the maximum lies within the second half of the window, the time shift is interpreted by DFT as negative and decreasing, and adds a positive and linearly increasing linear component to the phase spectrum. This linear increase is typical for all onsets and can be used for onset detection. Similarly, if a signal has an offset, a linearly increasing negative linear phase component is to be observed, which can be used for offset detection. If both onset and offset are present, i.e. most of the event is enclosed in the window, a discontinuous point will be observed where the positive and the negative components meet. To make these processes detectable, the step of sliding must be much smaller than the length of the window.

3. The Method

We base our proposed onset detector on the approach mentioned above. We denote the width of the window by N and its sliding step by N/M where $N > M >> 1$. At each slide t, the unwrapped phase spectrum $\varphi(t,k)$ is calculated and its linear component is extracted. Since the phase spectrum of any real signal is odd, this linear component is determined by only one parameter, the slope, which can be estimated from the first half of the phase spectrum by

$$\alpha(t) = \frac{24}{(N-2)(N-1)N}\sum_{n=0}^{N/2-1} n\varphi(t,n) \qquad (1)$$

The track of $\alpha(t)$ is then used for onset detection.

Remarks:

(a) In fact, only large values of $\alpha(t)$ make much sense in our task. A large value of $\alpha(t)$ means that $\varphi(t, k)$ makes many wraps modulo 2π, in which case $\alpha(t)$ is roughly equal to $2\varphi(t, N/2)/N$. Thus it's possible to use $\varphi(t, N/2)$ instead of $\alpha(t)$ for onset detection. In experiments this makes little difference to the performance. We call $N\alpha(t)/2$ the linear phase span, and $\varphi(t, N/2)$ the phase span, of the spectrum. We have $\varphi(t, N/2) \approx N\alpha(t)/2$ for large $\alpha(t)$'s.

(b) M shall be selected large enough to offer proper resolution within the width N. In practice we use M=4, 8 or 16.

(c) N shall be selected large enough so that the linear phase component becomes significant compared to other components. However, if the window is too wide, multiple maximums may be enclosed. This will confuse the detector, and make existing onsets bypassed.

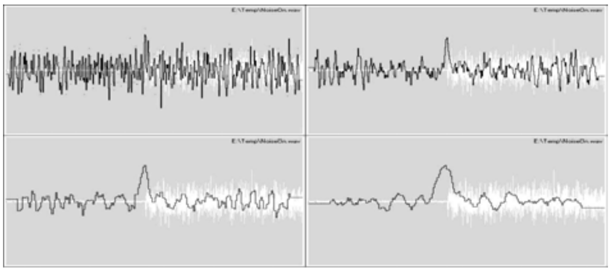

Figure 2. Phase span track

(a) N=64; (b) N=128;
(c) N=256; (d) N=512

Fig.2 shows the track curves of the phase span $\varphi(t, N/2)$ with respect to t, N ranging from 64 to 512. The first and second halves of the signal are white noises of different levels, with the second half having the higher level. An onset is to be detected where the curve reaches a maximum after being positive and almost linearly increasing for a duration of $N/2$. In this example N is required larger than 128 to separate the onset from noises.

A small problem with the method lies in phase calculation. As we've mentioned, an unwrapped phase angle calculated in close vicinities of zeroes could be meaningless and cause unwanted side effects. To ease this problem we multiply each phase jump between adjacent spectral samples by a weighting factor which favors "strong enough" samples equally and suppresses the contribution of very weak samples. In practice we select a threshold th and set the weights by

$$w_k = \begin{cases} c, & if\ H(k) > th\ and\ H(k-1) > th \\ c\,(\min[H(k), H(k-1)]/th)^2, & otherwise \end{cases}, \qquad (2)$$

where the kth weight w_k is applied to the phase jump $\varphi(k)$ - $\varphi(k\text{-}1)$, and c is a factor that keeps the sum of all weights equal to $N/2$-1.

Fig.3 shows the phase span track of an impulse signal. The phase spectrum of an impulse is strictly bilinear. Its phase span equals its linear phase span. The whole track is cut in half by the impulse. Before the impulse the track is positive and linearly increasing, featuring a pattern of onsets (Fig.3b); after the impulse the track is negative and linearly increasing, featuring a pattern of offsets (Fig.3c). The lengths of the onset and offset parts are $N/2$ each, and the heights of them are both $N\pi/2$. Comparing Fig.3b with Fig.2a through 2d,

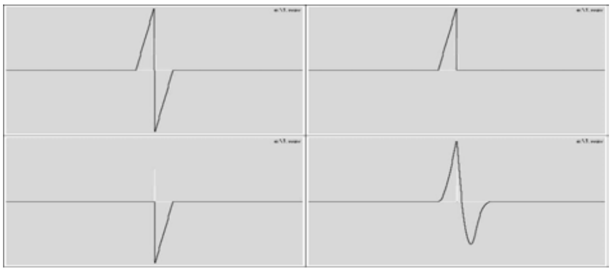

Figure 3. Phase span track of an impulse
(a) whole track; (b) onset part;
(c) offset part; (d) matched-filtered

it's easy to see how well the pattern in Fig.3b conforms to those in Fig.2. We'll use this pattern directly for onset detection. Likewise, the pattern in Fig.3c can be used for offset detection.

Pattern matching is performed by the conventional technique of matched filter, which is especially useful for fixed 1-D patterns like the one in Fig.3b. Given a fixed pattern $p(t)$ with spectrum $P(\omega)$, a matched filter is constructed by its impulse response $h(t)=kp(t_0-t)$, and its transfer function is given by $H(\omega)=kP^*(\omega)\exp(-j\omega t_0)$. Here k is a positive number, t_0 is a time shift to locate the maximum of the output, and * is the conjugate operator. For our pattern in Fig.3b, we set $t_0=N/2$, then according to

$$p(t)=\begin{cases} t\pi, & 0\le t<N/2 \\ 0, & otherwise \end{cases}, \tag{3}$$

we have

$$h(t)=\begin{cases} k(N/2-t)\pi, & 0\le t<N/2 \\ 0, & otherwise \end{cases}. \tag{4}$$

Therefore the output of the matched filter is written as

$$y(t)=k\sum_{t'=0}^{N/2-1}(\frac{N}{2}-t')\pi\varphi(t-t',\frac{N}{2}) \tag{5a}$$

when phase span is used as input, or as

$$y(t)=k\frac{N}{2}\sum_{t'=0}^{N/2-1}(\frac{N}{2}-t')\pi\alpha(t-t') \tag{5b}$$

when linear phase span is used as input. k is selected by

$$k=24/(N-2)(N-1)N\pi^2 \tag{6}$$

so that the maximum output of a perfect match is 1. Fig.3d depicts the matched-filtered result for an impulse.

4. Examples

In this section we give several examples on our phase-based onset detector. All samples are in 16kHz PCM.

4.1. White noise with level jump

Fig.4 depicts the output of the matched filter for the signal discussed in Fig.2. We see that the absolute level does not change the appearance of the curve, but a positive level jump signals an onset. This figure is obtained by setting N=256. The result remains unaltered as N ranges from 64 to 2048, as long as the level jump is above 7.5dB.

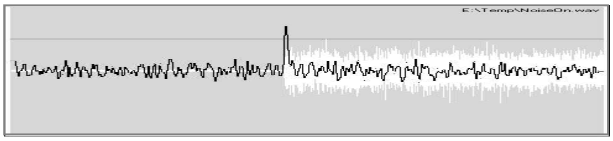

Figure 4. Level jump in white noise

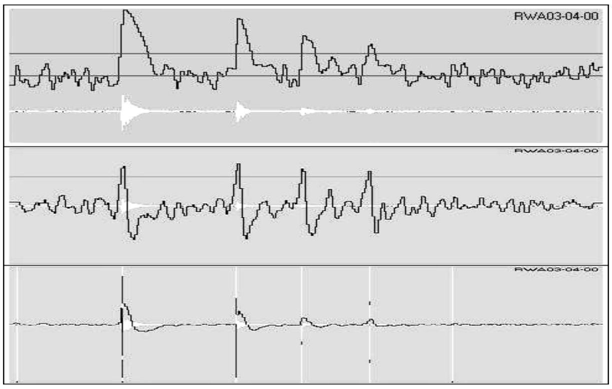

Figure 5. Sound of dropping a dice on wooden board
(a) the signal; (b) phase-based detector;
(c) multi-band amplitude-based detector

4.2. Sound of dropping a dice on a wooden board

This sample is from the RWCP sound scene database [6] indexed at RWA03-04-00. 4 impacts could be heard in the sample. Fig.5a shows the waveform (in white) and the track of its instantaneous sound level. Fig.5b gives the output of the matched-filter. For comparison we look at the result in Fig.5c of an amplitude-based onset detector using 33 frequency bins. In Fig.5c the vertical lines marks detected onsets and the curve describes the sharpness of the onsets. Each vertical line is composed of 33 white or dark segments. Each dark segment marks a frequency bin that evidences the existence of the onset.

In Fig.5c, while the leading impacts are detected in most bins and are shown strong, the others are detected only in a few bins and evaluated very weak. In Fig.5b the 4 peaks are almost the same height, stressing the existence rather than loudness of each onset. By setting the detection threshold at 0.5, the impacts are correctly detected. Again, this result remains unaltered when N ranges from 64 to 512. When N is larger than 1024, the sliding window encloses multiple events.

4.3. Door knock with rattles

This sample is from *Sound Ideas* sound effects CD 9011, at track 37 index 2. Its spectrogram is given is Fig.6a. Two major door knock impacts and four rattling events are heard and observed in the spectrogram. Comparison is made between our phase-based and amplitude-based detectors like in the previous example. Each of these rattles is detected in more than 6 bins by the amplitude-based detector (Fig.6c). The estimated sharpness distinguishes them better from the major impacts than from noises. The phase-based detector on the other hand, finds the rattles much closer to the major impacts than

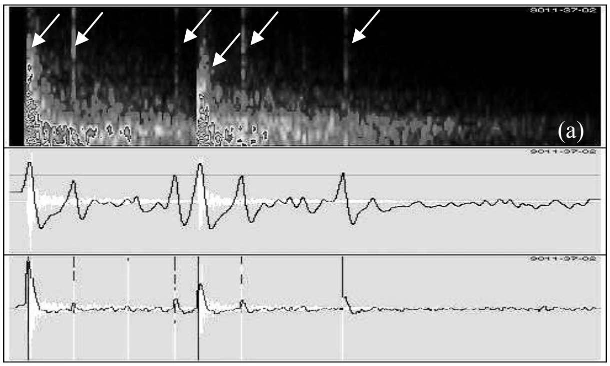

Figure 6. Sound of door knocks with rattles
(a) the spectrogram; (b) phase-based detector
(c) multi-band amplitude-based detector

to background noises, which again stresses the existence more than the strength of an onset.

4.4. Instrumental music

We select 3 instrumental music recordings. The first and second pieces are Bach's *Prelude in C, BWV 846a*, played by R. Kirkpatrick on the harpsichord and G. Gould on the piano, respectively. The third is Paganini's *Allegro vivo e spiritoso – Minore* from *Sonata No.6 for violin and guitar*, played by G. Shaham and G. Sollscher. All pieces are mono, down-sampled to 16kHz. Window width N=1024 is used. The positive peaks of $y(t)$ in (5a) are examined and counted in intervals $(0, 0.1]$, $(0.1,0.15]$, $(0.15,0.2]$, $(0.2,0.25]$ and $(0.25,1]$. Results are given in tables 1a through 1c.

Table 1a. Test on harpsichord

Interval	>0.25	0.2-0.25	0.15-0.2	0.1-0.15	<0.1	Total
Existing	431	86	28	4	0	549
Inserted	0	0	2	3		

Table 1b. Test on piano

Interval	>0.25	0.2-0.25	0.15-0.2	0.1-0.15	<0.1	Total
Existing	193	103	117	95	37	545
Inserted	15	3	8	15		

Table 1c. Test on violin

Interval	>0.25	0.2-0.25	0.15-0.2	0.1-0.15	<0.1	Total
Existing	431	53	57	38	27	606
Inserted	8	3	11	33		

The test on harpsichord yields the best result. We achieve zero detection miss and 0.91% wrong insertion at threshold 0.1, and 0.73% miss vs. 0.36% insertion at threshold 0.15. The piano piece has 545 notes. The miss and insertion rates are 6.79% vs. 7.52% at threshold 0.1 and 24.2% vs. 4.77% at threshold 0.15. The third piece is a duo of violin and guitar. Since the volume of violin is remarkably higher, only audible violin notes are studied. Those of guitar, if detected, are regarded as insertions. In this sense we have evaluated the miss and insertion rates as 4.46% vs. 9.08% at threshold 0.1 and 10.7% vs. 3.63% at threshold 0.15.

5. Discussion

Up to now we have relied our method on the pattern in Fig.3b, which is derived from a positive volume change. If an onset features only pitch or timbre transition but no amplitude increase, the method may fail.

Moreover, even if we forget pitch and timber, the pattern in Fig.3b is only necessary but never sufficient to assert an onset. According to Tab.1b, many insertions do have high peaks (>0.25). Close observation shows that most of these peaks are related to local inaudible discontinuities. Experiments show that we can create such high peaks artificially by adding a hardly audible offset to just one sample point in the waveform.

Despite all these, our experiments have proved the value of the simple method. A natural advantage of the method lies in the independence of phase on sound level, which enables the detection of very close onsets with very diverse intensities, requiring no adaptation to volume. Another advantage is that we are now able to use a long sliding window. Amplitude-based onset detectors have to use short windows to preserve timing accuracy, and become susceptible to noises and fast modulations. This does not trouble our phase-based detector since timing information is already encoded in the phase, so we're free to choose longer windows.

6. Conclusion

In this article we've described a simple but effective method to detect acoustical onsets using phase spectra only. The method is robust against volume changes, local noises and fast modulations. However, there are problems with the method, solutions of which rely on a more insightful decoding of the phase.

7. References

[1] Dolson M., "The Phase Vocoder: A Tutorial", *Comput. Music J. 10*, pp 14-27, 1986.

[2] Brown J., "A High-resolution Fundamental Frequency Determination Based on Phase Changes of the Fourier Transform", *J. Acoust. Soc. Amer.*, *Vol.94(2),* pp 662-667, 1993.

[3] Duxbury C., Sandler M., Davies M., "A Hybrid Approach to Musical Note Onset Detection", *Proc. DAFx-02,* pp 34-38, 2002.

[4] Klapuri A., "Sound Onset Detection by Applying Psychoacoustic Knowledge", *Proc. ICASSP99,* 1999.

[5] Smith L. S., "Onset-based Sound Segmentation", in Touretzky D.S., Mozer M.C., Hasselmo M.E. (eds), *Advances in Neural Information Processing Systems 8 (Proceedings of the 1995 Conference),* pp 729-735, MIT Press, 1996.

[6] Hiyane K., Iio J., "Non-speech Sound Recognition with Microphone Array", *Proc. HSC2001,* pp 107-110, 2001.