TIME DOMAIN NOTE AVERAGE ENERGY BASED MUSIC ONSET DETECTION

Ruolun Liu¹, Niall Griffith¹, Jacqueline Walker² and Peter Murphy²

¹: Department of Computer Science and Information System
²: Department of Electronic and Computer Engineering
University of Limerick, Limerick, Ireland

Ruolun.Liu@ul.ie

ABSTRACT

A novel time domain strategy is proposed for the detection of the onset of musical notes based on the changing energy level. By calculating the note average energy (NAE), the proposed method is insensitive to both the dynamic range of the energy levels in a piece and whether or not the piece is monophonic or polyphonic. More importantly, the new strategy tackles the thorny 'threshold' problem that is always being avoided unsuccessfully. The detection performance of the new method is illustrated by its performance over a range of music pieces played on different instruments.

1. INTRODUCTION

Onset detection is one of the most important components of computational auditory scene analysis (CASA). Subsequent analysis of rhythm, melody, and harmony depend on the accuracy of onset detection. It also of importance for transcription, music editing and recording.

Conventional onset detection methods are usually based on a peak-searching technique where a threshold is needed to detect onsets from a measured curve. The problem is how to determine the threshold. It is either superficially "properly chosen", or a fixed proportion of a specified value. From simulations, we found that the appropriate threshold varies with the music in question. A good threshold for piano doesn’t guarantee a good result for violin. Even worse, for the same piece of music, the threshold suitable in one part is not suitable for some other parts. None of the existing methods provides a generalized method for determining the threshold for music of different styles and instruments to avoid missing notes and spurious notes being detected.

The key issue for onset detection is how to determine the noticeable change in the sound energy. Since the noticeable level varies from song to song and from style to style, no fixed threshold is suitable for all kinds of music. However, is it possible to find the onsets without any threshold?

2. EXISTING METHODS

To try and answer this question, we investigate existing methods of measuring the prominence of the note energy level change. Among the existing methods, curves that can be used in the peak searching procedure are the different formulations of the evolution of sound energy. They can be roughly classified into time domain methods and frequency domain methods.

2.1. Time Domain Methods

To begin with, a power envelope must be extracted. Assume \( s(t) \) is the sound wave signal and \( h(t) \) is the smoothing window. The instant power envelope is the convolution of the instantaneous power, \( s'(t) \), and the smoothing window, that is:

\[
p(t) = \int_0^t s^2(\tau)h(t-\tau)d\tau
\]

While listening to music, people can relatively easily tell where the onset is, from the falls and rises of the instant power envelope \( p(t) \). To make a computer perform this task, a better curve is needed in which the peaks can be easily found.

First Order Difference & Evolution Ratio

Among the different developed formulations of \( p(t) \), the most traditional ones include First Order Difference and the Evolution Ratio,

\[
FOD(t) = p(t) - p(t-T) \quad (2)
\]

\[
R(t) = \frac{p(t)}{p(t-T)} \quad (3)
\]

where \( T \) is the sampling interval.

Both the first order difference and the evolution ratio measure the degree of change in the sound power level in a quite natural way. While \( FOD(t) \) observes the absolute change, \( R(t) \) measures the relative change. When the sound power increases, \( FOD(t) \) returns a positive value, and a negative value for decreasing power. On the other hand, the faster the power changes, the greater the degree by which \( R(t) \) differs from 1: greater than 1 for increase and less than 1 for decrease.

Since \( FOD(t) \) peaks at the highest local change in the power envelope, it doesn’t peak at the very beginning of the note. In addition, small oscillations in the decline of a note from a higher power level of \( p(t) \) may cause peaks that are larger than peaks corresponding to notes with relatively lower power levels. In Figure 1(b), the 4th note has 2 by-peaks that are bigger than the main peak for the 7th note.
Leaving aside this problem, $R(t)$ is better than $FOD(t)$. It peaks quite close to the beginning of the note, but if there are small sharp surges at lower power level in $p(t)$, it returns peaks competitive to the peaks corresponding to notes at a relatively higher level. As shown in Figure 1(c), after the 7th note, there are 2 extra peaks that are bigger than the main peak for the 5th note.

**First Order Relative Differential**

To overcome the problems of the first order difference, Klapuri\cite{1} proposed a first order relative differential which is the first order differential of $p(t)$ divided by $p(t)$ itself.

$$FORD(t) = \frac{p'(t)}{p(t)} \tag{4}$$

This is equivalent to the first order differential of the logarithm of $p(t)$. $FORD(t)$ is said to be psychologically relevant because people perceive the sound energy level in a logarithmical scale. Equal increases are more prominent in a quiet signal than in a loud signal. However, we found little difference between the first order relative differential and the evolution ratio, and it is hard to find any improvement, see Figure 1(d).

**2.2. Frequency Domain Methods**

The power envelope can also be obtained in the frequency domain from the short time Fourier transform (STFT) of the sound signal.

$$q(t) = \int_{0}^{+\infty} \left| STFT_s(f ; t) \right|^2 df \tag{5}$$

where $STFT_s(f ; t)$ is the STFT coefficient, at frequency $f$, of the sound frame of $s(t)$ beginning at time $t$. Two consecutive frames may also be overlapped.

Theoretically, the time domain methods mentioned above can also use the power envelope obtained by (5), which results in time-frequency hybrid methods; but practically, $q(t)$ given by (5) is not smooth enough to give good results. Researchers have developed frequency domain methods based on measurements from the frequency domain.

**High Frequency Content (HFC)**

$$HFC(t) = \int_{0}^{+\infty} f \left| STFT_s(f ; t) \right|^2 df \tag{6}$$

By weighting the power spectral intensity linearly toward the high frequency, HFC emphasises the high frequency component of the power spectral intensity. This strengthens the measurement of a sharp attack in a sound signal because a sharp attack involves more higher frequency components. However, $HFC(t)$ doesn’t peak at the start of the rise in $q(t)$. In Figure 2(b), it is easy to see $HFC(t)$ peaking at the middle of the rise of $q(t)$.

**Masri’s Detection Function**

Masri\cite{2} proposed a detection function based on $HFC(t)$ which peaks at the instant of the start of a new note.

$$M(t) = \frac{HFC(t)}{q(t)} \times \frac{HFC(t)}{HFC(t-T)} \tag{7}$$

$M(t)$ observes the relative power changes in both the higher frequency components and $HFC(t)$ for all frequency components. One big drawback of $M(t)$ is the imprecision of the peaks for notes with a slow rise. See Figure 2(c).

**FORD of HFC**

Giuliano\cite{3} proposed a combination of first order relative differential and high frequency content.
\[ D(t) = \frac{HFC'(t)}{HFC(t)} \]  

As shown in Figure 2(d), \( D(t) \) performs better than \( HFC(t) \) and Masri’s method to some degree, but the inherent shortcomings of the two original methods don’t automatically compensate for each other and no additional techniques are used to overcome them.

### 2.3. Band Wise Processing

Another possibility is to detect onsets \( o_i(t) \), within frequency bands selected by a Gammatone filter. By setting a weight to each band according to the contribution of the components in its band, we can expect a better detection of the onset.

\[ o_n(t) = \sum_{i=1}^{l} w_i o_i(t) \]  

This idea was first proposed by Scheirer\(^4\). It may be a good complement if no further improvement can be found without bandwise processing. But it’s also possible to create spurious onsets by frequency band splitting. We will now consider a solution that removes the threshold problem and is applicable to both monophonic and polyphonic music played on a wide range of instruments.

### 3. NOTE AVERAGE ENERGY ONSET DETECTION

As addressed in section one, the key issue for onset detection is how to determine ‘noticeable’ changes in sound energy. We noted that none of the methods outlined above really investigate the energy of the music signal. True, power reflects the energy quite precisely, but it’s not real energy after all.

Energy is the integration of power with respect to time.

\[ e(t) = \int_0^t p(t)dt \]  

In Figure 3(b), we can see a bending corresponding to each note, even some small sharp oscillations in its decline. However, the bending is not very clear because \( e(t) \) is an increasing function. As it is not convenient to find the start or end of the bending, we decided to investigate the average energy.

\[ a(t) = \frac{1}{l} \int_{t-l}^t p(t)dt \]  

In Fig.3(c), the bending is much easier to detect, but not for the later notes. Since the energy difference between consequent notes is bounded, but time \( t \) keeps increasing, when \( t \) is big enough, no energy difference could be found between two consequent notes from \( a(t) \).

Enlightened by the features of the averaged energy, \( a(t) \), we investigated \( a(t) \) within the duration of each note, namely Note Average Energy (NAE). We found as expected that NAE declines when no new note follows; while NAE rises when a following note adds additional energy. This suggested that NAE could be a better measurement of the ‘noticeability’ of energy level changes.

Suppose we know the onset instant, \( t_n \), of every note where \( n = 1, 2, \ldots, N \), and \( N \) is the number of notes. NAE is defined as:

\[ NAE(t) = \frac{1}{t-t_n} \int_{t_n}^t p(t)dt, \quad (t_n < t < t_{n+1}) \]  

When \( t = t_n \), \( NAE(t_n) = p(t_n) \). From Figure 3(d), \( NAE(t) \) has only one increase and decrease for each note. So onsets can be determined simply by checking the local minima.

Since \( \{ t_n \mid n = 1, 2, \ldots, N \} \) is the set of onsets to be determined, \( NAE(t) \) can not be calculated directly. However, this doesn’t mean we cannot decide where an onset is. Let \( n=0 \) and \( t=0 \), if we calculate \( NAE(t) \) in a range of \( t \) with a limit open to the right. If the following relation:

\[ NAE( t-T ) > NAE(t) < NAE(t+T) \]  

is met, we can set \( n = n+1 \) and \( t_n = t + T \), and move on to the next detection also in a range of \( t \) with a limit open to the right. In this way, we remove the threshold problem.

Other strengths of NAE include insensitivity to the overall sound energy variation, and applicability to both monophonic and polyphonic music. This comes from the inherent property of \( NAE(t) \). Since \( NAE(t) \) investigates one note only, the overall power evolution has little effect on onset detection. \( NAE(t) \) examines the onset from the view of note energy changing in the time domain, it has nothing to do with whether the music is monophonic or polyphonic. Since no frequency domain measurement is taken into account in the detection, the proposed method has no trade off between resolutions of time and frequency, and less complexity.

### 4. IMPLEMENTATION

To implement onset detection based on NAE, several additional techniques were used.

**Trimming:** keeps only instant power samples between the first and last two samples that are stronger than a certain proportion of the maximum. This procedure removes the
recording noise that causes extra spurious notes at the start and end.

*Time Warping:* compacts instant power samples of one frame by the maximum of this frame with a certain overlap. It can be treated as a special over-sampling that significantly reduces the size of data.

The trimmed, warped instant power signal is not smooth enough to be forwarded to detection. Considering that different music at different sampling rates has different smoothness, we introduce an *Adaptive Smoothing* technique, where the length of the smoothing window is decided according to the smoothness of curve to be smoothed. Smoothing is repeated until the number of peaks in the curve maintains the same for two consequent applications.

Though the overall power evolution has little affect on onset detection based on NAE, since NAE(t) is calculated with a right open range of t, a very strong note can still overshadow a weak following note. It will be better if we can lower the strong notes and raise the weak ones to the same level. *Evening:* the power envelope achieves this to some degree by dividing the smoothed power envelope by its baseline.

*Sharpening:* sets the decay of the evened, smoothed power envelope to be exponential. This gives a clearer change in NAE(t) between two consequent notes.

*Refining:* the detected onset time, \( t_n : n = 1, 2, \ldots, N \), can be refined by looking back into the smoothed power envelope to find a set of local minima. It can be refined still further by looking back into the instant power and repeatedly smoothing the corresponding frame of the instant power with the Hanning window of half of the frame length till the number of peaks in the curve is the same for two consecutive applications. With this baseline, the minimum, indicates time instant of the onset, \( t_n^\hat{} : n = 1, 2, \ldots, N \).

5. EXPERIMENTS

With a range of music pieces, we evaluated the proposed method using a measure of Correct Detection Rate (CDR), which is defined as:

\[
CDR = \frac{\text{total-missing-spurious}}{\text{total}} \times 100\% \quad (14)
\]

“total” takes the bigger of the number of true notes and number of notes detected. The music covers different instruments and genres and dynamic ranges. The first six songs are generated from a midi score, so we have the actual onset for comparison. The rest are taken from the recordings of symphonies and pop song, so the timing cannot be evaluated. Cut-off for trimming, warping frame length and frame overlap are 1/20, 1ms and 50% respectively for all the songs. From the results we obtain, the CDR is very satisfactory.

6. CONCLUSIONS

Based on the observation that existing note onset detection methods detect the onset from the real energy of sound, we developed a new detection scheme using the NAE. NAE focuses on the energy within one note, and thus it relatively insensitive to both the dynamic range of the song and whether the song in question is monophonic or polyphonic. The new strategy removes the threshold problem with a better detection performance for a large range of music pieces of different instruments. The proposed method has no trade off between time resolution and frequency resolution.

7. REFERENCES


<table>
<thead>
<tr>
<th>Song</th>
<th>True</th>
<th>Detected</th>
<th>Missing</th>
<th>Spurious</th>
<th>CDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bass</td>
<td>15</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>93.75%</td>
</tr>
<tr>
<td>Piano</td>
<td>32</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Trombone</td>
<td>22</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Ocarina</td>
<td>52</td>
<td>52</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Banjo</td>
<td>52</td>
<td>52</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Clarinet</td>
<td>52</td>
<td>52</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Violin (vivacissimo)</td>
<td>809</td>
<td>832</td>
<td>12</td>
<td>35</td>
<td>94.35%</td>
</tr>
<tr>
<td>Ava Maria (Bach)</td>
<td>813</td>
<td>907</td>
<td>23</td>
<td>71</td>
<td>89.64%</td>
</tr>
<tr>
<td>Canon (Pachedel)</td>
<td>556</td>
<td>825</td>
<td>21</td>
<td>248</td>
<td>67.39%</td>
</tr>
<tr>
<td>Rikki (Blues)</td>
<td>68</td>
<td>71</td>
<td>3</td>
<td>6</td>
<td>87.32%</td>
</tr>
</tbody>
</table>

Table 1: Summary of detection result based on NAE.