Wavelet Analysis for Onset Detection

Crawford Tait (craw@dcs.gla.ac.uk), William Findlay (wl@dcs.gla.ac.uk)
Department of Computing Science, University of Glasgow, G12 8QQ

Abstract

The first stage of many higher level analysis, detection and synchronisation tasks is the identification of note onsets. We present a method that involves no prior knowledge of the input signal, and may be applied to non-musical sounds. The technique highlights note onsets by various ensembles in the preceint of overlapping notes, and will also attempt to specify which harmonics have begun as the onset. A semitone-based wavelet analysis is used to generate a time-frequency plane of modulus values, that is then transformed according to metrics derived from the study of auditory perception. The plane is then viewed as a series of vectors, and calculation of the distance between groups of vectors adjacent in time shows peaks in the distance function at onset locations. The final stage of interpretation involves detecting peaks in this function, and classifying the peaks as onsets, or otherwise.

We show a selection of examples, and describe the results of an experiment conducted to investigate the effects of loudness envelope and harmonic complexity.

1. Background

Onset detection has mostly been attempted in the context of automatic transcription ([Poulet et al 83], for example). Because of this, and since amplitude information alone is insufficient in all but the most simple examples, early attempts often relied on pitch detection. However, pitch detection is potentially a more complex problem, and there are many situations in which detecting the onset of timbres would be useful.

An alternative approach utilises knowledge of the particular instruments involved, however this makes generalisation difficult ([Goto and Murakami 93]).

A technique tailored to detecting bass and snare drums, for example.

All methods share some time-frequency decomposition as a first stage, and [Smith 93] describes how a decomposition based on the human auditory system might be used in conjunction with a neural network to locate onsets in individual frequency bands. Whilst this does not rely on pitch detection, it appears difficult to automate the interpretation of the onset data (an analogous situation is described herein).

Before explaining the decomposition we have adopted, it should be noted that, to date, almost all onset detection attempts have been based on the assumption of monophonic input. This means that only a single melodic line is present, and notes do not overlap. Whilst this is also the case here, the method we will describe is capable of locating onsets in the presence both of interfering notes, and the environmental reverberation present on most recordings.

2. Wavelet Analysis

Traditionally, time-frequency decomposition of audio signals has been achieved via Fourier analysis. However, its linear division of the frequency scale does not correspond well with our perception of pitch (which is logarithmically related to frequency). In addition, the basis sinusoids are not localized in time – audio is usually split into short time segments, and an average spectrum is calculated for each on the (usually) assumption of its periodicity.

These problems have prompted the investigation of wavelet analysis. The basis functions are well localized in both frequency and time, and there is an inherent logarithmic division of the frequency scale. A time series of Gabor coefficients is generated for each frequency band (or scale level), with time resolution dictated by bandwidth and centre frequency of the band. From these, planes of modulus or phase values can be calculated (an introduction in the context of signal processing is given in [Koel & Vetterli 91]).

Wavelet analyses have, however, mainly been based on octave frequency bands, which are too wide in a musical context.

Early investigations showed that discontinuities in the input signal produced convergent lines of constant phase ([Kronland-Martinet et al 87]). Work on an onset detection method based on this observation (which also uses semitone division of the frequency range) is described in [Solbach et al 95]. It is unclear, though, how this method would cope with a range of examples, including those with reverberation (which can disrupt the phase plane).

Our analysis is based solely on the modulus plane, and uses the harmonic wavelet analysis of [Newland 95]. This is efficient, and allows a semitone division of the frequency scale, making it ideally suited to musical input (for main area of concern). We will show that the unavoidable decrease in time resolution is not, in practice, a handicap.

Figures 2.1 and 2.2 show the examples we will use throughout this paper. The computed modulus values are mapped to dot densities and plotted against time (upper and lower scale levels containing negligible energy have been omitted).
3. Transforming the Modulus Plane

Whilst we do not aim to create an auditory model, observations of the auditory system have informed further transformations of the raw modulus plane. For example, perception of loudness is related to amplitude on a logarithmic scale, so that there is some justification for mapping the modulus values to a logarithmic scale. In fact, this procedure does enhance significant features, as well as improving the results of further analyses.

Also, the auditory system is not equally sensitive to all frequencies. This implies that a weighting could be applied to each level of scale, based on its centre frequency, to emphasise features that occur in frequency ranges where auditory sensitivity is highest. Many researchers have attempted to derive such a measure experimentally. (Stevens 73 presents a frequency weighting function derived by combining the results of a large number of such studies, and an approximation of this (extrapolated to the low and high frequency regions not covered in the study) has been applied to the output of the wavelet transform.

The last transformation we have implemented is adaptive normalisation of the modulus values. This is necessary, because the quieter notes is a passage tend to be dwarfed by the louder notes. We overcome this by first finding the maximum modulus values in adjacent time windows, and then interpolating. A normalisation and thresholding factor is then found which adapts to the current loudness. This is not directly related to the auditory system, but overcomes the problem presented by dynamic variation in the input (as may be observed in figure 2.1, which becomes quieter towards the end).

In summary, the raw modulus plane is first adaptively normalised, and then mapped to a logarithmic scale.

The transformed versions of the modulus planes in figures 2.1 and 2.2 can be seen in figures 5.1 and 5.3.

4. Highlighting Onsets

In order to investigate time varying behaviour, the plane of modulus values is divided into vectors, each constituting a slice through the plane at the highest time resolution. Vectors are compared by treating them as points in $N$-dimensional space (if $N$ senstive bands are being analysed), and considering the distance between them as given by the Euclidean norm:

$$d_{ij} = \sqrt{\sum_{k=0}^{N-1} (M_{jk} - M_{ik})^2}$$

where $M_{ik}$ is the modulus value for sensitive $k$ in the vector at time $i$, and sensitive levels 0 to $N^2$ at times $i$ and $j$ are being considered. A vector distance will thus exist if a change in frequency and/or amplitude takes place between the two points in time under consideration.

The method we have adopted is to calculate the distance between the average of the vectors in two adjacent sliding windows (whose size is dictated by a minimum note length). The a-erages are representative of the state of the modulus plane over a short interval, and peaks in the calculated function are evident even when only gradual change occurs between notes. Of course, the output of this method is somewhat smoothed, and the peaks are broadened — but this makes automatic detection of (the peaks much easier, and is acceptable as long as short notes are not obscured. In fact, a further aid peak detection, the output is smoothed by replacing each calculated distance by an average of itself and a few others to either side.

The results of applying this technique to the transformed versions of figures 2.1 and 2.2 are shown in figures 4.1 and 4.2.

Both of these examples are from orchestral recordings, with considerable reverberation, and are
in legato style (with onsets not heavily punctuated).
However, peaks are produced at all onset locations.

The peaks are detected by simply locating points at
which a series of successive increases is followed by
a series of successive decreases. A small threshold
is also introduced, and peaks must be separated by
at least the minimum note length.

Many peaks will correspond to onsets, however
others will correspond to offsets and some will be
spurious. Therefore, we must attempt to classify each
peak based on the behaviour of the modulus plane
around its location in time.

Currently, this is achieved by traversing scale levels
at the location of the onset, and locating the point in
the analysis window at which the difference between
the averages on either side is maximized. A level is
marked as a partial onset if its average is increasing
and its later average is both above a threshold, and
significant when compared with the change between
the two averages. Partial onsets are counted and a
peak is classified as an onset if several are detected.

Figures 5.1 to 5.4 show the results of applying this
procedure (with identical parameters) to the examples
already illustrated.

Fig. 5.1 - detected onsets marked on modulus plane of
clarinet piece.

Fig. 5.2 - detected onsets marked on amplitude envelope of
clarinet piece.

Every onset except the last is located in the first
example, and the second shows all onsets marked
(with a single spurious detection, third from last).

Fig. 5.3 - detected onsets marked on modulus plane of
French horn piece.

6. Conclusions and Further Work

The harmonic wavelet transform provides a useful
tool for analysis of all kinds of sound, but more work
is required in its practical application (one of the
authors has described a vector-based similarity
measure for detecting repetition in [Tait 95]). We feel
that the onset detection method described herein
performs well on a variety of cases (in that onsets
generally produce peaks in the distance function),
however the classification of onsets could be
improved.

Also, a persistent problem in the course of this
research has been the lack of a set of established test
cases. In [Tait & Findlay 95], we presented an
experiment to investigate the effects of loudness
evelope and harmonic complexity. This showed
that, even in the presence of overlapping notes, onsets
involving slow attacks produced peaks in the vector
distance function. However, there are many more
variables and a set of recordings providing some
degree of coverage is required.

We have begun to design an experiment which would
be based around a single piece of music (including
various phrases and dynamics). This would be
baked around a single piece of music (including
various phrases and dynamics). This piece would then
be played on a range of instruments which
would be intended to cover different types of attack
characteristics and timbres. In addition, results
obtained in the presence of glissando, degrees of
variance and tremolo degrees of vibrato and
fretboard, and some representative non-musical
sounds would have to be examined separately.

Finally, we illustrate the benefit of not relying on
pitch perception, which widens the scope of possible
application considerably. Figures 6.1 to 6.3 show the
results of analysing the sound of 11 footsteps, with
background music increasing in loudness (the regular
rhythm is interrupted by a scraping floor between
steps eight and ten).

In examples with greater polyphony, or background
sounds, adaptive normalisation is not to appropriate
and only logarithmic scaling has been applied in
figure 6.1 (the distance in figure 6.3 were calculated
exactly as before, however the classification
parame-ters were changed slightly in order to
highlight all of the steps).
Fig 6.1 - modulus plane of footsteps with background music, logarithmic scaling applied (JTO 88).

Fig 6.3 - footstep marked in the amplitude envelope.

Fig 6.3 - distances from figure 6.1.

References


(Goto and Murakoa 95) - "Music Understanding at the Beat Level - Real-time Beat Tracking for Audio Signals", Manataka Goto and Yoichi Murakoa; ICAI 95 Computational Auditory Scene Analysis workshop.

(JTO 88) - "Koel's Korney" from "Wait A Minute": The James Taylor Quartet; Polydor CD 837 342-2.


[Mozart 74] - Concerto for Clarinet and Orchestra in A major, K. 622 (Adagio); Polydor CD 432 352-2.


[Schubert 94] - Symphony No.9 in C (Great, Andante - Allegro ma non troppo); BBC CD MM117.

[Smith 95] - "Using an Oost-based Representation for Sound Segmentation": Leslie S. Smith; Submitted to NEURAP70.

[Solbach et al 95] - "The Complex-valued Conjugate Wavelet Transform as a Preprocessor for Auditory Scene Analysis": Ludger Solbach, Rolf Kohrmann & Jorg Kleeberg; ICAI95 Computational Auditory Scene Analysis workshop.


ICMC Proceedings 1996 503