A HYBRID APPROACH TO MUSICAL NOTE ONSET DETECTION

Chris Duxbury, Mark Sandler, Mike Davies

DSP and Multimedia Group,
Dept. of Electronic Engineering,
Queen Mary, University of London
christopher.duxbury@elec.qmul.ac.uk

ABSTRACT

Common problems with current methods of musical note onset detection are detection of fast passages of musical audio, detection of all onsets within a passage with a strong dynamic range and detection of onsets of varying types, such as multi-instrumental music. We present a method that uses a subband decomposition approach to onset detection. An energy-based detector is used on the upper subbands to detect strong transient events. This yields precision in the time resolution of the onsets, but does not detect softer or weaker onsets. A frequency based distance measure is formulated for use with the lower subbands, improving detection accuracy of softer onsets.

We also present a method for improving the detection function, by using a smoothed difference metric. Finally, we show that the detection threshold may be set automatically from analysis of the statistics of the detection function, with results comparable in most places to manual setting of thresholds.

1. BACKGROUND

Note onset detection aims to find the start of musical events from the audio signal itself. It is an essential component of many larger systems such as automatic musical transcription schemes, non-linear time scaling [1], and many new audio effects and editing tools, such as 'beat detective'[2] from Digidesign. It is also common for many synthesis applications to require isolation of the attack portions of notes.

Despite some proposed solutions, it remains an unsolved, and often over-simplified, problem. Traditional methods such as high frequency detection rely on the assumption that all note onsets contain high frequency energy [3]. The assumption that, for most instruments, a note will contain more high frequency energy at its onset is fair to make. However, in the case of real world audio examples where there may be high notes with considerable high frequency energy at their onset in the same region as low notes with weak high frequency energy, the lower notes become almost impossible to detect from the detection function. This work addresses this problem directly.

If we consider the nature of musical signals, there is a range of different types of instrument onsets. Figure 1 shows short sections of signals from a guitar and a violin. The guitar is a string instrument that is played percussively, leading to 'hard' note onsets, appearing as wide-band noise in the spectogram. For this type of instrument, high frequency content is a useful detection method. However, the violin in this figure is an example of a bowed string instrument, with a 'soft' onset. The strings are excited because of the stick-slip caused by the friction of the bow. In this case, the notes are being excited constantly, hence there is little, or no, decay. Here, the change in frequency content, particularly at lower frequencies, is our best guide to note onsets. Most everyday musical signals contain a range of hard and soft onsets.

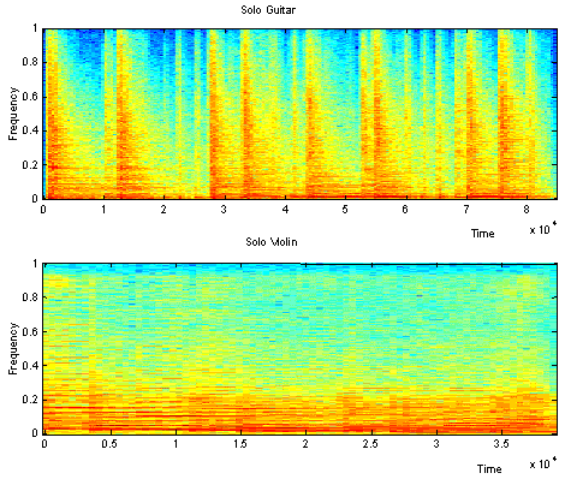

Figure 1: *Spectogram of solo guitar (upper) and solo violin (lower) signals. The guitar has a hard percussive attack, whereas violin onsets have a much softer onset.*

To overcome this, several proposals use energy content in individual frequency bands to search for possible onsets [4],[5]. This improves results; however using the same energy content algorithm across all frequency bands is not necessarily the best method. This suggests low notes would be just as easily detected as high notes, which is not the case.

Consider high frequencies: there are often short bursts of energy with large gaps between, relative to frequency. At lower frequencies, the notes decay more slowly, and occur at a faster rate relative to frequency, suggesting basic energy detection will not be as effective.

There are some 'grey area' issues with note onset detection that require consideration. Notes which fade-in, rather than have a hard onset, are problematical for most methods of onset detection. Likewise, glissando (smooth transitions between notes) can lead to cases of wrong detection. Both these issues are improved by the multiresolution hybrid scheme offered here, with more slowly occurring onsets appearing in the longer-windowed lower subband.

It is also worth noting that there are some differences between

note onset detection and transient detection, such as [6],[7]. The former is concerned with detecting the beginning of musical events, whilst the latter aims to isolate fast changes. Although the results produced may be similar, as there are usually fast changes at note onsets, transient detection alone will not detect notes with softer onsets, such as slower attacks.

Our aim is to produce a note onset detection scheme that yields good results for the full range of musical instruments and attacks, regardless of signal, at a low computational cost. Further to this, the detector should require no user inputs such as manual threshold setting or information on the signal or instrument type.

For this reason we propose using a hybrid scheme of transient energy detection in the high frequency subbands, with an FFT-based distance measure used to detect note changes at low frequencies. This facilitates detection of both the hard and soft onsets shown in figure 1.

2. SUBBAND HYBRID DETECTION SCHEME

The signal is split into a number of frequency subbands for individual onset detection analysis. This is implemented using a constant-Q conjugate quadrature filter bank, as described in [8], with 5 bands from 0-1.1kHz up to 11-22kHz. From individual analysis of each subband, it is clear that the highest band contains weak onset information for almost all signals. To save computation this band ($> 11kHz$) is not used. The next three subbands representing the range from 1.2-11 kHz contain noticeable bursts of energy for a range of note onsets. The lowest subband does not have the same strong bursts of energy at note onsets - however, there are noticeable differences in the frequency content at note changes. From this, we propose using standard energy content analysis only for the upper subbands (1.2-11kHz). This yields excellent detection only for those signals with wideband-noise based onsets, and also yields results which are accurately localized in time. The subband energy, $SE(n)$, is given by:

$$SE(n) = \sum_{m=(n-1)h}^{nh} |x(m)|^2 \quad (1)$$

where m is the time index, n is the hop number and h is the hop size. h may be short($\approx$ 128 samples) as the downsampling of the subband scheme means it varies in each band. This effectively yields the temporal envelope of the signal, sub-sampled by a factor h relative to the samplerate of the subband. In the upper subbands, note onsets can be detected from jumps in energy, using the difference:

$$ons(n) = SE(n) - SE(n-1) \quad (2)$$

This has the advantage that it is computationally efficient, whilst yielding good results for a range of signals. In section 3 we show how this idea can be extended further using a one sided smoothing function. However, at a cost of an additional short-time Fourier transform (STFT) in each upper subband, we can utilize a transient energy measure [9]. Considering basic phase vocoder principles, it is expected that, for steady state frequency components, the instantaneous frequency should be approximately equal in adjacent frames. The transient energy, TE(n), is therefore given by:

$$TE(n) = \sum_{k \in K_{tr}} |X(k, nh)|^2 \quad (3)$$

where K_{tr} is used to denote the set of transient frequency bins, k_{tr}, given by:

$$\phi(k_{tr}, (n-2)h) - 2\phi(k_{tr}, (n-1)h) + \phi(k_{tr}, nh) < T_{tr} \quad (4)$$

The subband transient energy term then replaces the subband energy terms in equation (2). This eliminates any upper steady state components such as high frequency partials, or high pitched notes, but at a much greater computational cost. Implementations using both subband energy and transient subband energy have been tested, and the transient energy approach is found to eliminate some of the detection function noise in recordings with greater high frequency content. However, in the upper subbands this improvement is not significant, and to reduce computation may only be used for the mid-frequency content.

For the lower two frequency bands (0-2.5kHz, representing the range of musical notes up to $D\#_7$) we propose the use of a distance measure between the vectors created for each frame of an FFT. This is based on a standard Euclidean distance measure, EDM:

$$EDM = \sum_{k=1}^{N/2} \{|X(k, nh)| - |X(k, (n-1)h)|\}^2 \quad (5)$$

However, it is clear from equation (5) that fast decay will have the same effect as a note onset. Whilst this may be desirable for location of attack transients, as in [3] where a similar distance measure is used, our aim is to locate note onsets for musical analysis, editing and effects. Hence, we take only positive values. Putting:

$$dX_n(k) = X(k, nh) - X(k, (n-1)h) \quad (6)$$

we then have the distance measure:

$$DM = \sum_{\{k; dX_n(k) > 0\}} dX_n(k)^2 \quad (7)$$

A normalization term is incorporated so that softer onsets are detected alongside harder onsets:

$$DM = \frac{\sum_{k; dX_n(k) > 0} dX_n(k)^2}{\sum_{k=1}^{N/2} |X(k, (n-1)h)|^2} \quad (8)$$

Frequency domain smoothing improves results by limiting effects of instabilities. This produces a detection function that varies from a standard energy measure in that it looks at the average over one frame of the increases in energy content within each FFT track. In the lower frequency band, this detects all changes in note, but at a cost of poorer time resolution. Hence, each method is a trade-off between correct detection and good time localization of detected onsets.

At higher frequencies, the measured frequencies of the partials are not stable enough for using a distance measure, as illustrated by the minimal improvement offered by the transient energy function. Note that there is some frequency domain overlap (in the 1.2-2.5 KHz subband) between the methods, increasing data for analysis, however, this allows for the region where there is still noticeable energy bursts, but also slowly decaying pitch information.

A key advantage to using a subband filterbank of this type is the good time resolution in the upper subbands, to locate hard onsets, whilst there is good frequency resolution at low frequencies, making a distance measure more useful. It would be desirable to tie this to psychoacoustic principles as the ear acts as a filterbank,

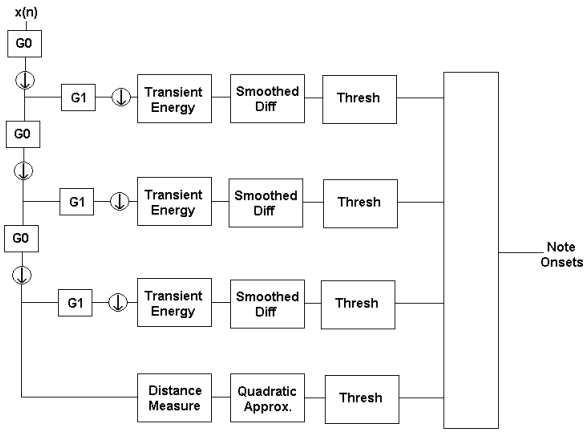

Figure 2: *Block Diagram of Proposed Onset Detection Scheme. GO and G1 represent the low pass and high pass CQF filters respectively.*

however, it has been shown [10] that the time localisation of sound onsets is not frequency dependant above 200Hz, but appears to be linked to the bandwidth of the signal at the onset.

A block diagram of the overall hybrid subband scheme is shown in figure 2.

3. DETECTION FUNCTIONS

The detection function from each subband now takes the form of an energy/distance measure over time. Within this signal, we are interested in the sharp increases. For this reason, the derivative of the energy/distance measure we have calculated in the previous section would typically be used. However, immediately after an onset, there are often noisy regions leading to multiple detections. One common solution is to low pass filter the energy/distance, however, this leads to a blurring of the position of onsets, as well as a smoothing of weaker onsets.

We propose using the difference between the current frame, and several previous frames. Adapting equation (2) this detection function is now given by:

$$ons(n) = \sum_{a=1}^{A} \frac{SE(n) - SE(n-a)}{W(a)} \tag{9}$$

where $W(a)$ is a weighting function of the integer a.

This gives a meaningful interpretation in the time domain, however, by re-arranging, we see that this produces a detection function that is based on the difference between the signal, and a smoothed version of itself:

$$ons(n) = K.SE(n) - \sum_{a=1}^{A} W^{-1}(a)SE(n-A) \tag{10}$$

where

$$K = \sum_{a=1}^{A} W^{-1}(a) \tag{11}$$

Note that K is a constant term outside the summation, and may therefore be ignored. From this it is clear that the weighting function $W^{-1}(a)$ acts as filter coefficients of a filter. If no weighting function is used, this is the equivalent of low pass filtering with a fast transition band, which increases with the number of coefficients, A. This clearly gives too much weighting to energy terms which occur a long time before the onset. The weighting terms, $W^{-1}(a)$ tested were linear and exponential. In the linear case of:

$$ons(n) = SE(n) - \sum_{a=1}^{A}(1 - \frac{a}{A})SE(n-a) \tag{12}$$

Compared to the case with no weighting function, this gives a greater weighting to the subband energy terms before the potential onset location. In filtering terms, this is a low pass filter with smoother roll off than the previous case.

This was compared to the exponential weighting function given by:

$$ons(n) = SE(n) - \sum_{a=1}^{A} \frac{SE(n-a)}{a} \tag{13}$$

Exponential weighting now gives much greater emphasis to more recent energy values, whilst allowing previous values to have some effect. This function produces the clearest detection function results. In filtering terms, the signal is low pass filtered with a very smooth roll-off, as most of the energy is in the first filter coefficients. This also reduces the effect of the choice of A ($A = 30$ in our implementation).

This approach to detection function smoothing maintains better time-localization of the onsets than basic low pass filtering, whilst minimizing multiple detections of onsets (see Figure 3).

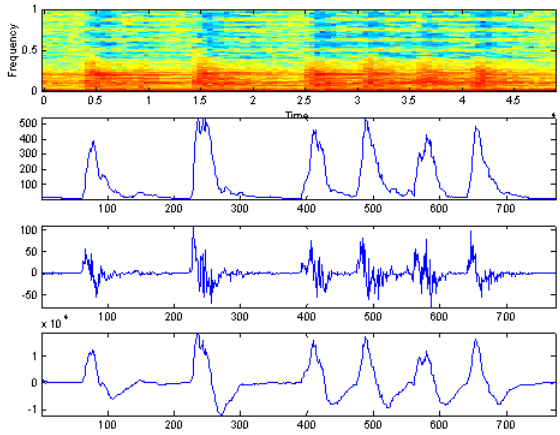

Figure 3: *Guitar signal spectogram (top) with 5.5kHz-11kHz subband energy measure over time (upper middle). Traditional derivative-based detection function (lower middle), compared to proposed detection function (bottom).*

4. AUTOMATIC THRESHOLD SETTING

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 loss of weaker transients, and poorer time resolution. 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 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. By using the statistical properties of the detection function, we propose a method for automatic setting of a threshold. Figure 4 shows a typical histogram of the detection function described in the previous section.

The onsets are defined as the outliers within the histogram, whereas the no-onsets content should be closer to zero. The detection function histogram may be viewed as a combination of two probability density functions:

$$p(tr) \sim N(0, \sigma_{tr}^2) \tag{14}$$

$$p(nt) \sim N(0, \sigma_{nt}^2) \tag{15}$$

where $p(tr)$ is probability of a transient, and $p(nt)$ is probability of not a transient. Where transients are present in the signal, it is expected that:

$$\sigma_{tr} >> \sigma_{nt} \tag{16}$$

The no-onset probability function will have a high peak at its approximately zero mean, with a narrow distribution. Conversely, the onset detection probability distribution will have a low peak at zero, representing those cases where onsets produce a low detection function amplitude, with a wide distribution. The combination of these produces a histogram like that of figure 4.

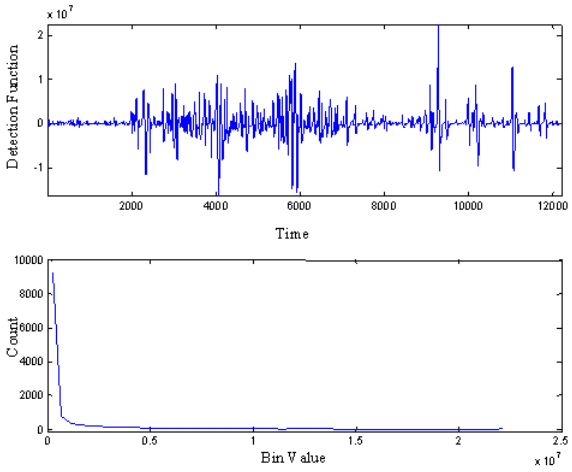

Figure 4: *One-sided histogram (lower) of 2.25-5.5kHz subband detection function of audio signal (upper).*

The ideal threshold is therefore at the point where the data is more likely to be an onset. We proposed two methods allowing this. The first of these used a mixture of two Gaussians, fitted using the EM algorithm, as described in [[11]]. However, this approach proved costly in terms of computation when compared with our second approach. The second uses the second derivative of the histogram as an approximation of the first method in order to reduce computation.

The aim is to find the point where all greater values in the detection function represent note onsets. If we study the proposed model shown in figure 5, the threshold should be set at the position where the combined pdf curve takes the characteristic of the transient component. This occurs at the maximum of the second derivative.

The threshold can be set in this manner for any size window of data. However, due to the variety of content within a signal, windows of approximately 5 seconds are used in this scheme.

We intend to extend the scheme to use the statistics of the signal to find regions where onsets are not present within a subband, such that the onset probability should be zero. This may offer a solution to some of the problems outlined in the results section.

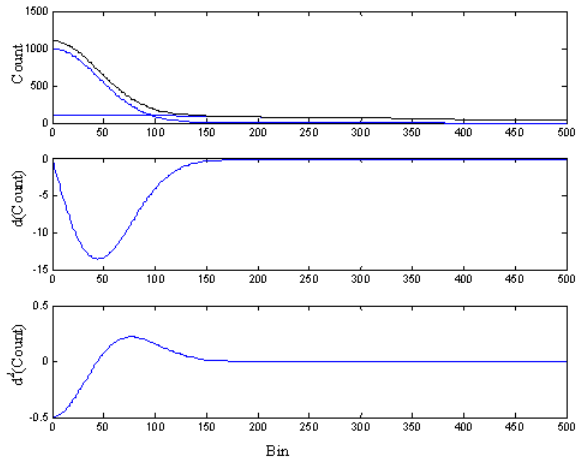

Figure 5: *Approximate model of pdfs for transients(with wide variance) and non-transient by fitting two Gaussians (top), with their total shown as a dotted line. The first derivative is shown below, with the second derivative as the bottom plot. The threshold is set at the maximum in the second derivative.*

5. COMBINING SUBBAND INFORMATION

Subband information may be combined and onsets detected in the combined function such as:

$$S_{all} = S_1(t) + S_2(t) + S_3(t) + S_4(t) \tag{17}$$

If peak picking is applied to this function, it does not solve the problem of weaker onsets remaining undetected, even though they may be strong within a certain frequency band. For this reason, we choose to detect onsets from the subband detection functions, and combine the results.

After peak-picking and thresholding each subband detection function, we have a range of detected onsets. Each output of the subband scheme produces positions of onsets. In many cases, onsets are present in multiple, although not all, subbands. If all these are taken, we obtain:

$$P(t) = P_{S1}(t) + P_{S2}(t) + P_{S3}(t) + P_{S4}(t) \tag{18}$$

where $P(t)$ is a signal length vector containing one at onsets and zeros elsewhere, $S1$ denotes the highest subband, and $P_{Sx}(t)$ is a signal length zero vector containing one at onsets detected in subband x and zeros elsewhere.

However, many onsets will appear in more than one subband, with some difference in position cause by the resolution differences between subbands and the intrinsic differences between the two methods of the hybrid scheme. For this reason, we take a short window of 50ms and take a maximum of one onset for each window. All other onsets within this window are then discarded.

This approach is adopted rather than position averaging as time resolution is improved in the upper subbands, whilst detection accuracy is improved in the lower subband in this scheme. As the higher frequency subbands have the best time resolution, a higher band always takes precedence over a lower band, optimizing the results obtained. This is done by weighting the output of each subband:

$$P(t) = \alpha P_{S1}(t) + \beta P_{S2}(t) + \gamma P_{S3}(t) + P_{S4}(t) \qquad (19)$$

where α, β, and γ are weighting terms such that:

$$\alpha > \beta > \gamma \qquad (20)$$

and $\gamma > 1$. $P(t)$ is windowed such that only the greatest weight onset is kept within the $50ms$ window. The remaining onsets are discarded.

A second reason for using a weighting scheme of this nature is that it may be tuned so that only 'hard' or 'soft' onsets are selected. In [1] we presented a time-scaling algorithm which required phase to be locked at hard onsets. Here it is essential that all note onsets are detected, whilst hard and soft onsets are treated separately. This onset weighting scheme has been successfully applied to the time-scaling algorithm. Hard onsets can be defined as those with a greater value for $P(t)$ within the $50ms$ window before any subband onsets are discarded. This can be inferred because hard onsets should appear as strong across several subbands. It is documented that human listeners detect transients more easily as their bandwidth increases [10].

6. RESULTS

The proposed onset detection scheme was tested with a range of signals. A variety of instruments, as well as performance and musical styles were tested. For each signal tested, onset points were assigned by a listener beforehand. Due to the time consuming nature of this, short musical segments (approx. 12 seconds) were used in these tests.

An onset is considered accurately detected if the measured onset falls within 50ms of the pre-determined onset position. An onset is undetected if no onset is measured within the 50ms window. If an onset is detected outside the 50ms window around a pre-determined onset position, it is considered a false detection. From [4], the measure of onset accuracy used is given by:

$$Accuracy = \frac{N_t - N_{ud} - N_{fd}}{N_t}.100 \qquad (21)$$

where Nt represents total number of actual onsets in the signal, N_{ud} represents number of onsets undetected by algorithm, and N_{fd} represents number of false detections.

For each signal, the test was run twice. In the first case, the algorithm was tested with automatic threshold setting, and no user setting of parameters. This represents the black box case that is required for an onset detection scheme to be incorporated into larger systems. In the second case, the same algorithm was tested with user defined thresholds for each of the subbands.

The signals which were used to test the algorithm were:
$Jazz1$ - Solo Jazz Guitar signal containing both single notes and chords played with a wide range of dynamics.
$Jazz2$ - Solo Jazz Guitar signal containing fast passages of notes.
$Dido$ - Pop music example with voice, guitar, keyboards, bass and drums.
$Piano1$, Piano2- Solo piano signals containing both single notes and chords played with a wide range of dynamics
$Opera$ - Classical signal containing only bowed strings.

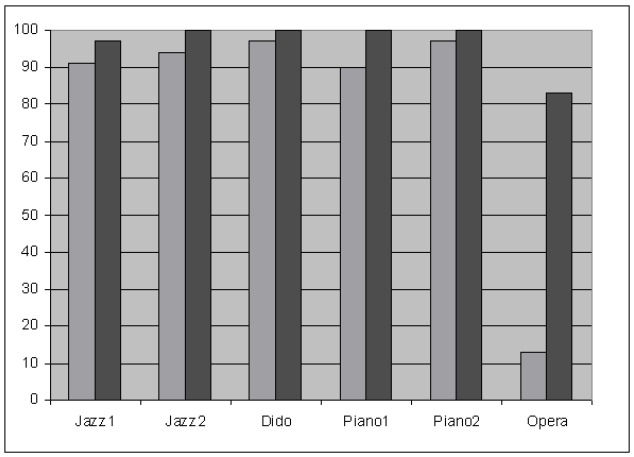

Figure 6: *Results of note onset detection for a range of signals. For each signal results are shown with both automatic threshold setting (light grey) and manual threshold setting (dark grey).*

As shown in figure 6, this approach to onset detection was found to yield high quality results with above 90 percent accuracy for a range of musical signals. In particular it performed equally well for both hard and soft note onsets, across the whole frequency spectrum. The results showed considerable improvements on standard energy methods in the detection of softer, low notes, improving missed detection rates with little increase in false alarms. Very fast legato passages cause the most errors, however, this is also the case with human perception.

The key problem shown in these results is for the automatic threshold setting for those signals where onset information is weak in certain subbands. In this case, the automatic threshold is set low, leading to massive over-detection. This problem is illustrated by the Opera test signal results. By increasing the threshold from the automatic threshold to a high threshold in the upper subbands, the results increase from 13% to 84% accuracy. We have looked at ways to overcome this problem such that subbands are ignored if they appear to contain weak onset information. This is essentially a measure of whether the signal appears to be purely noisy, or contain some outlier information, as explained in section 4. However, this is intended for further investigation in future work.

7. CONCLUSIONS

A musical onset detection scheme has been proposed which takes advantage of subband decomposition, a hybrid approach to detection, an improved detection function approach and automatic setting of detection thresholds. It goes some way to solving the problem of whether energy or frequency content approaches should be used exclusively, by applying each to the frequency bands within which they are most relevant.

The algorithm was found to yield good results for a wide range of musical signals, with no threshold or parameter setting requirements. Further improvements were made with human input, however this kind of approach is unrealistic for most applications requiring note onset detection.

It may be the case that other note onset or transient detection schemes, such as the statistical approach of [12] could offer yet further improvements, and a deeper investigation into other approaches is intended as future work. The idea of a scheme that offers 100% detection rate for all signals is still a long way off. Whether it will ever be possible is a matter for discussion. However, assigning note onsets accurately is often a difficult and time consuming task for human listeners.

8. ACKNOWLEDGEMENTS

Thanks to Laurent Daudet for his comments and suggestions.

9. REFERENCES

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

[2] Digidesign, "Pro tools 5.1.1 software specifications," http://media.digidesign.com/products/docs/prd_1057_2041.pdf.

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

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

[5] F. Jaillet X. Rodet, "Detection and modeling of fast attack transients," in *Proc. Int. Comp. Music Conf. (ICMC,'01)*, 2001.

[6] L. Daudet, S. Molla, and B. Torresani, "Transient detection and encoding using wavelet coefficient trees," in *Proc. of the GRETSI'01 conference*, 2001.

[7] T.S. Verma, S. Levine, and T.H.Y. Meng, "Transient Modeling Synthesis: A flexible analysis/synthesis tool for transient signals," 1997.

[8] A. Haddad, Ed., *Multiresolution Signal Decomposition*, Academic Press, 1992.

[9] C. Duxbury, M. Davies, and M. Sandler, "Extraction of Transient Content in Musical Audio using Multiresolution Analysis Techniques," in *Proc. Digital Audio Effects Conference (DAFX,'01)*, 2001.

[10] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, fourth edition, 1997.

[11] C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, 1995.

[12] Thornburg H. and Gouyon F., "A Flexible Analysis-Synthesis Method for Transients," in *Proc. of the International Computer Music Conference (ICMC2000)*, 2000.